Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics

Preslav Ivanov Nakov

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-173
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-173.html

December 20, 2007

Acknowledgement

First, I would like to thank Marti Hearst for being such a fantastic advisor: for helping and encouraging me throughout my graduate studies, and for giving me the freedom to pursue my ideas, but also for pushing me to get things done in time when it was necessary. Next, I would like to thank all members of my dissertation committee, Marti Hearst, Dan Klein, Jerome Feldman, and Lynn Nichols for their comments, which were invaluable for shaping the final version of the present thesis. Thanks to all present and former members of the BioText group at Berkeley: I will really miss you guys! I am especially grateful to Ariel Schwartz, Anna Divoli, and Barbara Rosario for the many fruitful discussions and collaborations throughout the years.

Using the Web as an Implicit Training Set:
Application to Noun Compound Syntax and Semantics

by

Preslav Ivanov Nakov

Magistar (Sofia University, Bulgaria) 2001

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Marti Hearst, Co-chair
Professor Dan Klein, Co-chair
Professor Jerome Feldman
Professor Lynn Nichols

Fall 2007

The dissertation of Preslav Ivanov Nakov is approved:

Professor Marti Hearst, Co-chair Date

Professor Dan Klein, Co-chair Date

Professor Jerome Feldman Date

Professor Lynn Nichols Date

University of California, Berkeley

Fall 2007

Using the Web as an Implicit Training Set:

Application to Noun Compound Syntax and Semantics

Abstract

Using the Web as an Implicit Training Set:

Application to Noun Compound Syntax and Semantics

by

Preslav Ivanov Nakov

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Marti Hearst, Co-chair

Professor Dan Klein, Co-chair

An important characteristic of English written text is the abundance of noun compounds – sequences of nouns acting as a single noun, e.g., *colon cancer tumor suppressor protein*. While eventually mastered by domain experts, their interpretation poses a major challenge for automated analysis. Understanding noun compounds' syntax and semantics is important for many natural language applications, including question answering, machine translation, information retrieval, and information extraction. For example, a question answering system might need to know whether '*protein acting as a tumor suppressor*' is an acceptable paraphrase of the noun compound *tumor suppressor protein*, and an information extraction system might need to decide if the terms *neck vein thrombosis* and *neck thrombosis* can possibly co-refer when used in the same document. Similarly, a phrase-based machine translation system facing the unknown phrase *WTO Geneva headquarters*, could benefit from being able to paraphrase it as *Geneva headquarters of the WTO* or *WTO headquarters located in Geneva*. Given a query like *migraine treatment*, an information retrieval system could use paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

I address the problem of noun compounds syntax by means of novel, highly accurate unsupervised and lightly supervised algorithms using the Web as a corpus and search engines as interfaces to that corpus. Traditionally the Web has been viewed as a source of page hit counts, used as an estimate for n-gram word frequencies. I extend this approach by introducing novel surface features and paraphrases, which yield state-of-the-art results for the task of noun compound bracketing. I also show how these kinds of features can be applied to other structural ambiguity problems, like prepositional phrase attachment and noun phrase coordination. I address noun compound semantics by automatically generating paraphrasing verbs and prepositions that make explicit the hidden semantic relations between the nouns in a noun compound. I also demonstrate how these paraphrasing verbs can be used to solve various relational similarity problems, and how paraphrasing noun compounds can improve machine translation.

Professor Marti Hearst
Dissertation Committee Co-chair

Professor Dan Klein
Dissertation Committee Co-chair

To Petya.

Contents

Acknowledgments

First, I would like to thank Marti Hearst for being such a fantastic advisor: for helping and encouraging me throughout my graduate studies, and for giving me the freedom to pursue my ideas, but also for pushing me to get things done in time when it was necessary.

Next, I would like to thank all members of my dissertation committee, Marti Hearst, Dan Klein, Jerome Feldman, and Lynn Nichols for their comments, which were invaluable for shaping the final version of the present thesis. Many additional comments on particular thesis topics are due to Roxana Girju, Elena Paskaleva, Zornitza Kozareva and Ayla Algar. I am also very grateful to Michael Jordan for having served in the committee for my qualifying exam, and to Galia Angelova, my master's thesis advisor, who was the first to introduce me to the field of computational linguistics.

Thanks to all present and former members of the BioText group at Berkeley: I will really miss you guys! I am especially grateful to Ariel Schwartz, Anna Divoli, and Barbara Rosario for the many fruitful discussions and collaborations throughout the years. Many thanks to all other (former) members of the group: Gaurav Bhalotia, Janice Hamer, Emilia Stoica, Mike Wooldridge, Brian Wolf, Jerry Ye, Harendra Guturu, Rowena Luk, Tingting Zhang, Alex Ksikes and Megan Richardson.

Last but not least, I would like to express my deepest gratitude for the love and support of my family – to my wonderful wife Petya, to my mother Ivanka, and to my brother Svetlin – for their support and encouragement throughout the long years of my graduate studies.

Chapter 1

Introduction

"*Recent studies identify the colon cancer tumor suppressor protein adenomatous polyposis coli (APC) as a core organizer of excitatory nicotinic synapses.*"

`http://www.tufts.edu/sackler/tcvr/overview.htm`

An important characteristic of technical literature is the abundance of long sequences of nouns acting as a single noun, which are known as noun compounds. While eventually mastered by domain experts, noun compound interpretation poses major challenge for automated analysis. For example, what is the internal syntactic structure of *colon cancer tumor suppressor protein*: '[*colon cancer*] [[*tumor suppressor*] *protein*]' or '[[*colon cancer*] [*tumor suppressor*]] *protein*' or '[[[*colon cancer*] *tumor*] *suppressor*] *protein*', etc.? Can *colon cancer* be paraphrased as '*cancer that occurs in the column*'? Or as '*cancer in the column*'? What is the relationship between *colon cancer* and *tumor suppressor protein*? Between *colon* and *cancer*? Is a *tumor suppressor protein* a kind/type of *tumor suppressor*? Is it a kind of *suppressor*?

Understanding noun compounds' syntax and semantics is important for many natural language applications including question answering, machine translation, information retrieval, and information extraction. A question answering system might need to know whether '*protein acting as a tumor suppressor*' is an acceptable paraphrase of the noun compound *tumor suppressor protein*, and an information extraction system might need to decide if the terms *neck vein thrombosis* and *neck thrombosis* could possibly co-refer when used in the same document. Similarly, a phrase-based machine translation system facing the unknown phrase *WTO Geneva headquarters* could benefit from being able to paraphrase it as *Geneva headquarters of the WTO* or *WTO headquarters located in Geneva*. Given a query like *migraine treatment*, an information retrieval system could use suitable paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

This gives rise to the following problems:

- **Syntax**: What is the internal syntactic structure of the noun compound? E.g., is it [[*neck vein*] *thrombosis*] or [*neck* [*vein thrombosis*]]?
- **Paraphrasing**: Which verbs can paraphrase the noun compound? E.g., '*thrombosis that <u>blocks</u> the neck vein*' is a good paraphrase for *neck vein thrombosis*.
- **Semantics**: What are the semantics of the noun compounds?
- **Semantic entailment**: Dropping which nouns yields a hypernym of the noun compound? E.g., is *neck vein thrombosis* a kind of *neck thrombosis*? See Appendix A.

I focus on the first three problems, which I address with novel, highly accurate algorithms using the Web as a corpus and search engines as interfaces to that corpus.

Traditionally in natural language processing (NLP) research, the Web has been used as a source of page hit counts, which are used as an estimate for n-gram word frequencies. I extend this approach with novel surface features and paraphrases which go beyond simple n-gram counts and prove highly effective, yielding a statistically significant improvement over the previous state-of-the-art results for noun compound bracketing.

I further address noun compound semantics by automatically generating paraphrasing verbs and prepositions that make explicit the hidden semantic relations between the nouns in a noun compound. Using them as features in a classifier, I demonstrate state-of-the-art results on various relational similarity problems: mapping noun-modifier pairs to abstract relations like `TIME`, `LOCATION` and `CONTAINER`, classifying relation between nominals and solving SAT verbal analogy questions.

Finally, I demonstrate the potential of these techniques by applying them to machine translation, prepositional phrase attachment, and noun phrase coordination.

The remainder of the thesis is organized as follows:

- In chapter 2, I discuss noun compounds from a linguistic point of view: First, I describe the process of compounding in general, as a mechanism of producing a new word by putting two or more existing ones together. Then, I focus on the special case of *noun compounds*: I discuss their properties, I provide several definitions (including one of my own to be used throughout the rest of the thesis), and I describe some interesting variations across language families. Finally, I explore some important theories, including Levi's (1978) theory of recoverably deletable predicates, Lauer's (1995) prepositional semantics, and Rosario *et al.*'s (2002) descent of hierarchy.

- In chapter 3, I describe a novel, highly accurate lightly supervised method for making left vs. right bracketing decisions for three-word noun compounds. For example, [[*tumor suppressor*] *protein*] is left-bracketed, while [*world* [*food production*]] is right-bracketed. Traditionally, the problem has been addressed using unigram and bigram frequency estimates to compute adjacency- and dependency-based word association scores. I extend this approach by introducing novel surface features and paraphrases extracted from the Web. Using combinations of these features, I demonstrate state-of-the-art results on two separate collections, one consisting of terms drawn from encyclopedia text, and another one extracted from bioscience text.

- In chapter 4, I present a novel, simple, unsupervised method for characterizing the semantic relations that hold between nouns in noun-noun compounds. The main idea is to look for *predicates* that make explicit the hidden relations between the nouns. This is accomplished by writing Web search engine queries that restate the noun compound as a relative clause containing a wildcard character to be filled in with a verb. A comparison to results from the literature and to human-generated verb paraphrases suggests this is a promising approach. Using these verbs as features in classifiers, I demonstrate state-of-the-art results on various relational similarity problems: mapping noun-modifier pairs to abstract relations like `TIME` and `LOCATION`, classifying relation between nominals, and solving SAT verbal analogy problems.

- Chapter 5 describes an application of the methods developed in chapters 3 and 4 for noun compound paraphrasing to an important real-world task: *machine translation*. I propose a novel monolingual paraphrasing method based on syntactic transformations at the NP-level, which augments the training data with nearly equivalent sentence-level syntactic paraphrases of the original corpus, focused on the noun compounds. The idea is to recursively generate sentence variants where noun compounds are paraphrased using suitable prepositions, and vice-versa – NPs with an internal PP-attachment are turned into noun compounds. The evaluation results show an improvement equivalent to 33%-50% of that of doubling the amount of training data.

- Chapter 6 describes applications of the methods developed in chapter 3 to two important structural ambiguity problems a syntactic parser faces – *prepositional phrase attachment* and *NP coordination*. Using word-association scores, Web-derived surface features and paraphrases, I achieve results that are on par with the state-of-the-art.

- Web search engines provide easy access for NLP researchers to world's biggest corpus, but this does not come without drawbacks. In chapter 7, I point to some problems and limitations of using search engine page hits as a proxy for n-gram frequency estimates. I further describe a study on the stability of such estimates across search engines and over time, as well as on the impact of using word inflections and of limiting the queries to English pages. Using the task of noun compound bracketing and 14 different n-gram based models, I illustrate that while sometimes causing sizable fluctuations, variability's impact generally is not statistically significant.

- Finally, chapter 8 lists my contributions and points to directions for future work.

I believe these efforts constitute a significant step towards the goal of automatic interpretation of English noun compounds.

Chapter 2

Noun Compounds

In this chapter, I discuss noun compounds from a linguistic point of view: First, I describe the process of compounding in general, as a mechanism of producing a new word by putting two or more existing ones together. Then, I focus on the special case of *noun compounds*: I discuss their properties, I provide several definitions (including one of my own to be used throughout the rest of the thesis), and I describe some interesting variations across language families. Finally, I explore some important linguistic theories, including Levi's (1978) theory of recoverably deletable predicates, Lauer's (1995) prepositional semantics, and Rosario *et al.*'s (2002) descent of hierarchy.

2.1 The Process of Compounding

The *Dictionary of Grammatical Terms in Linguistics* defines the process of *compounding* as follows (Trask 1993):

> "The process of forming a word by combining two or more existing words: *newspaper*, *paper-thin*, *babysit*, *video game*."

As the provided examples show, the words forming an English compound can appear orthographically separated, connected with a hyphen, or concatenated; often the same compound can be written in different ways, e.g., *health care*, *health-care*, *healthcare*.

Since the process of compounding constructs new words, these words can in turn combine with other words to form longer compounds, and this process can be repeated indefinitely, e.g., *orange juice*, *orange juice company*, *orange juice company homepage*, *orange juice company homepage logo*, *orange juice company homepage logo update*, etc.

Some authors impose further restrictions, e.g., Liberman & Sproat (1992) consider 'true' compounds only the ones representing lexical objects. Under their definition, *honeymoon* would be a compound, but *honey production* would not be.

Chomsky & Halle (1991) give a phonological definition: the words preceding a noun form a *compound* with it if they receive the primary stress. Therefore, *blackboard* is a compound (*black* gets the primary stress), but *black board* is not (equal stress). This definition is appealing since it stems from a standard linguistic test of whether a sequence of lexemes represents a single word, and is consistent with Trask's (1993) definition above, which views the process of compounding as a new word formation process. However, it is problematic since pronunciation can vary across different speakers. It is also of limited use for most computational approaches to language analysis, which work with written text, where information about the stress is not available.

On the other hand, knowing whether a sequence of words represents a compound might be useful for a speech synthesis system, where using the wrong stress can convey a meaning that is different from what is intended. For example, using a fronted stress would make *French teacher* a compound with the meaning of *'teacher who teaches French'*, while a double stress would convey the non-compound meaning of '*teacher who is French*' (Levi 1978). For compounds longer than two words, the correct pronunciation also depends on their internal syntactic structure, which makes a noun compound parser an indispensable component of an "ideal" speech synthesis system. If the internal syntactic structure is known, the stress assignment algorithm of Chomsky *et al.* (1956) can be used, which follows the constituent structure from the inside out, making the stress of the first constituent primary and reducing the rest by one degree. For example, if *black board* is not a compound, each word would get a primary stress [*black/1*] [*board/1*], but if it is a compound then the stress of the second word would be reduced: [*black/1 board/2*]. Now, if [*black/1 board/2*] is a sub-constituent of a longer compound, e.g., *black board eraser*, the rule would be applied one more time, yielding [[*black/1 board/3*] *eraser/2*], etc.

Although I am interested only in noun compounds, in English the process of compounding can put together words belonging to various parts of speech: noun+noun (e.g., *silkworm*, *honey bee*, *bee honey*, *stem cell*), adjective+noun (e.g., *hot dog*, *shortlist*, *white collar*, *highlife*), verb+noun (e.g., *pickpocket*, *cutthroat*, *know-nothing*), preposition+noun (e.g., *counter-attack*, *underwater*, *indoor*, *upper-class*), noun+adjective (e.g., *trigger-happy*, *army strong*, *bulletproof*, *dog tired*, *waterfull*, *English-specific*, *brand-new*), adjective+adjective (e.g., *dark-green*, *south-west*, *dry-clean*, *leftmost*), verb+adjective (e.g., *timbledown*), preposition+adjective (e.g., *overeager*, *over-ripe*), noun+verb (e.g., *hand wash*, *finger-point*, *taperecord*), adjective+verb (e.g., *highlight*, *broadcast*, *quick-freeze*), verb+verb (e.g., *freeze-dry*), preposition+verb (e.g., *overestimate*, *withdraw*, *upgrade*, *withhold*), noun+preposition (e.g., *love-in*, *timeout*), *breakup*), verb+preposition (e.g., *countdown*, *stand-by*, *cut-off*, *cast-away*), adjective+preposition (e.g., *blackout*, *forthwith*), preposition+preposition (e.g., *within*, *without*, *into*, *onto*). In a typical compound, the second word is the head, and the first one is a modifier, which modifies or attributes a property to the head. The part of speech of the compound is the same as that of the head. Some authors also allow complex structures like *state-of-the-art*, *part-of-speech* or *over-the-counter eye drop*.

2.2 Defining the Notion of *Noun Compound*

There is little agreement in the research community on how to define the notion of *noun compound*.[1] Different authors have different definitions, focusing on different aspects, and often use different terms in order to emphasize particular distinctions. Lauer (1995) provides the following list of closely related notions, used in the literature: *complex nominal*, *compound*, *compound nominal*, *compound noun*, *nominal compound*, *nominalization*, *noun compound*, *noun premodifier*, *noun sequence*, *noun-noun compound*, *noun+noun compound*. While some of these terms are broader and some are narrower in scope, most of them refer to objects that are syntactically analyzable as nouns (Chomsky & Halle 1991; Jackendoff 1975). Some of their building elements however can belong to other parts-of-speech, e.g., *complex nominals* allow for adjectival modifiers (see below).

Since noun compounds are syntactically analyzed as nouns, and since most of them exhibit some degree of lexicalization, it might be tempting to make them all lexical entries. Unfortunately, since the process of compounding can theoretically produce an

[1]Under most definitions, the term *noun compound* itself is an example of a noun compound.

unlimited number of noun compounds, this would effectively open the door to an infinite lexicon, which would create more problems than it would solve. Levi (1978) proposes the following solution for English (which does not apply to all kinds of noun compounds though): "*Complex nominals are all derived from an underlying NP structure containing a head noun and a full S in either a relative clause or NP complement construction; on the surface, however, the complex nominal is dominated by a node label of N.*"

Downing (1977) defines a *nominal compound* as a sequence of nouns which function as a single noun, e.g., *orange juice*, *company tax policy*, *law enforcement officer*, *colon cancer tumor suppressor protein*, etc. While being more restrictive than most alternatives, this definition is both simple and relatively unambiguous, which makes it the preferred choice for computational purposes.

Quirk *et al.* (1985) allow for any constituent to form a *premodified noun* with the following noun, e.g., *out-of-the-box solution*. Unfortunately, such a liberal definition makes it hard to distinguish a noun premodifier from an adjectival modification.

Levi (1978) introduces the concept of *complex nominals*, which groups three partially overlapping classes: *nominal compounds* (e.g., *doghouse*, *deficiency disease*, *apple cake*), *nominalizations* (e.g., *American attack*, *presidential refusal*, *dream analysis*) and *nonpredicate NPs*[2] (e.g., *electric shock*, *electrical engineering*, *musical critism*).

None of the above definitions can clearly distinguish between noun compounds and other nominal phrases. Levi (1978) lists the following three basic criteria that have been proposed for this purpose in the literature: *fronted stress*, *permanent aspect*, and *semantic*

[2]Nonpredicate NPs contain a modifying adjective which cannot be used in predicate position with the same meaning. For example, the NP *solar generator* could not be paraphrased as *'*generator which is solar*'.

specialization. While each of them captures some useful characteristics that are valid for most noun compounds, they are of limited use as tests for distinguishing a noun compound from other nominals.

The first criterion (that I have already introduced above), states that noun compounds should have a fronted stress. For example, *baby photographer* would be a noun compound only if pronounced with a fronted stress, (in which case it means ‘photographer who photographs babies’), but not with a normal stress (in which case it would would mean ‘photographer who is a baby’). This criterion is problematic since stress could differ across dialects and even across speakers of the same dialect. It also separates semantically parallel examples, e.g., it accepts *apple cake* as a noun compound, but not *apple pie*, which is undesirable.

The second criterion requires that the words forming a noun compound be in a permanent or at least habitual relationship, e.g., *desert rat* can only refer to rats that are strongly associated with a desert, e.g., living in or around it. Unfortunately, this criterion rejects examples like *heart attack*, *car accident*, and *birth trauma*.

The last criterion asks that noun compounds be at least partially lexicalized. Unfortunately, lexicalization is a matter of personal judgement and interpretation. For example, while *birdbrain* is an accepted lexicalization with a specialized meaning, ‘*ham sandwich*’ (referring to a person ordering a ham sandwich) is an innovative one with a more general meaning, many variations on which can be constructed on the fly. Lexicalization is also a matter of degree which makes it problematic to use as a test. Some noun compounds are completely lexicalized and therefore would be considered single lexical item/single lexical entries, e.g., *honeymoon* has nothing to do with *honey* or a *moon*. Other could be argued to be productively composed, e.g., *orange peel*. Many other lie in the continuum between, e.g., *boy friend* and *healthcare* exhibit a low degree of lexicalization. Towards the other end lie the metaphorical *ladyfinger* and *birdbrain*, which are highly lexicalized, but are still partially transparent, e.g., *ladyfinger* is a pastry that resembles a lady finger, and *birdbrain* is a person whose brain supposedly has the size of a bird's brain.

The partial lexicalization can sometimes be seen by the variation in spelling, e.g., both *health care* and *healthcare* are commonly used, and the latter seems to suggest the author's belief of a higher degree of lexicalization compared to the space-separated version. Similarly, the concatenated *bathroom* is more lexicalized than *game room*. In some cases, a high degree of lexicalization can be signalled by spelling changes in the compounded form, e.g., *dough* + *nut* = *donut*.

While most of the theories of the syntax and semantics of noun compounds focus primarily on nonlexicalized or lightly lexicalized noun compounds, for the purposes of my study, I do not try to distinguish or to treat differently lexicalized vs. nonlexicalized noun compounds. Note however, that my definition for a noun compound in section 2.4 does not allow for word concatenations, which readily eliminates most of the highly lexicalized noun compounds.

Lexicalization and transparency are related, but different notions. Taking for example the days of the week, *Sunday* is highly lexicalized, but is quite transparent: one can easily see '*sun* + *day*'. *Friday* is just as highly lexicalized, but is much less transparent in English. For comparison, the German for Friday, *Freitag* looks quite transparent '*frei*

+ *Tag'*, i.e., '*free day*', but this interpretation is actually wrong: both *Friday* and *Freitag* are historically derived from a translation of the Latin *dies Veneris*, which means '*day of Venus*'. Words that are borrowed as readily-constructed compounds would often be opaque to English speakers, e.g., *hippopotamus* comes from the Greek *hippopotamos*, which is an alteration of *hippos potamios*, i.e., '*riverine horse*'.

Noun compounds are sub-divided with respect to transparency into *endocentric* and *exocentric* ones, defined by the *Lexicon of Linguistics*[3] as follows:

> *Endocentric compound*: a type of compound in which one member functions as the head and the other as its modifier, attributing a property to the head. The relation between the members of an endocentric compound can be schematized as '*AB* is (a) *B*'. Example: the English compound *steamboat* as compared with *boat* is a modified, expanded version of boat with its range of usage restricted, so that *steamboat* will be found in basically the same semantic contexts as the noun *boat*. The compound also retains the primary syntactic features of *boat*, since both are nouns. Hence, a *steamboat* is a particular type of *boat*, where the class of *steamboats* is a subclass of the class of *boats*.
>
> *Exocentric compound*: a term used to refer to a particular type of compound, viz. compounds that lack a head. Often these compounds refer to pejorative properties of human beings. A Dutch compound such as *wijsneus* '*wise guy*' (*lit.* '*wise-nose*') (in normal usage) does not refer to a nose that is wise. In fact, it does not even refer to a nose, but to a human being with a particular property. An alternative term used for compounds such as *wijsneus* is *bahuvrihi* compound.

In general, English noun compounds are right-headed, but this is not always the case, e.g., *vitamin D* and *interferon alpha* are left-headed, and some noun compounds like *sofa-bed*, *coach-player* and *programmer analyst* are headless. The latter are known as coordinative, copulative or *dvandva* compounds; they combine nouns with similar meaning, and the resulting compound may be a generalization rather than a specialization. English also allows for reduplication, e.g., *house house*.

[3] `http://www2.let.uu.nl/Uil-OTS/Lexicon/`

Finally, for noun compounds of length three or longer, there can be multiple readings due to structural ambiguity. For example, *plastic water bottle* is ambiguous between a left- and a right-bracketing:

(1) [[*plastic water*] *bottle*] (left bracketing)
(2) [*plastic* [*water bottle*]] (right bracketing)

The correct interpretation is (2), with the meaning of a '*water bottle that is made of plastic*'; in (1) we have a *bottle* that has something to do with "*plastic water*" (which does not exist). As I mentioned in section 2.1 above, in spoken English, most of the time, the correct interpretation will be signalled by the stress pattern used by the speaker. Another way to signal the structure at speech time is by putting a pause at the appropriate position: '*plastic water* __ *bottle*' vs. '*plastic* __ *water bottle*'.

As I will show below, other languages can have more accurate mechanisms for resolving such structural ambiguities.

2.3 Noun Compounds in Other Languages

In this section, I briefly describe some interesting characteristics of noun compounds in other languages and language families, including Germanic languages, whose noun compounds are written without separating blanks. Romance languages, whose noun phrase (NP) structure is predominantly left-headed, Slavic languages (Russian, which forms noun compounds using various grammatical cases, and Bulgarian, which respects the original head-modifier order when borrowing foreign noun compounds), and Turkic languages, which mark the noun compound head with a special possessive suffix.

2.3.1 Germanic Languages

As I mentioned above, noun compounds in English are typically written with blank separators between the words, except for short and established lexicalizations like *textbook*, *newspaper*, *homework*, *housewife*, *Sunday*. In some cases, hyphenated forms are used, mainly for *dvandva* compounds, e.g., *coach-player*, *member-state*, and in compounds of length three or longer, e.g., *law-enforcement officer*. In the other Germanic languages, however, noun compounds are almost exclusively concatenated, e.g., *nagellackborttagningsmedel* (Swedish, '*nail polish remover*'), *sannsynlighetsmaksimeringsestimator* (Norwegian, '*maximum likelihood estimator*'), *kvindehåndboldlandsholdet* (Danish, '*the female handball national team*'), *Sprachgruppe* (German, '*language group*'), *wapenstilstandsonderhandeling* (Dutch, '*ceasefire negotiation*'). Concatenated noun compounds are pronounced with a fronted stress, which is lost if the words are written separately. Therefore, concatenation is very important; for example, in Norwegian the concatenated front-stressed form *smørbrød* means '*sandwich*', while the normal stress form *smør brød* means '*butter bread*'. This is similar to the way stress changes meaning in English; see the *English teacher* example in section 2.1 above.

Because of the concatenated spelling, noun compounds can be the source of very long words. The classic example in German is *Donaudampfschiffahrtsgesellschaftskapitän*, which stands for '*Danube steamship company captain*'. However, since there are no theoretical limits on noun compounds' length, even longer ones have been coined and used, e.g., *Donaudampfschiffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft*, meaning '*association of subordinate officials of the head office management of the Danube steamboat electrical services*', which is the name of a pre-war club in Vienna.

2.3.2 Romance Languages

Unlike in Germanic languages, in Romance languages noun phrases are generally left-headed, i.e., the modifier generally follows the head it modifies, e.g., *urânio enriquecido* (Portuguese, lit. '*uranium enriched*', i.e. '*enriched uranium*'). The same principle naturally extends from adjectives-noun modification to noun compounds, e.g., *estado miembro* (Spanish, lit. '*state member*', meaning '*member-state*'), or *legge quadro* (Italian, lit. '*law framework*', meaning '*framework law*'). In some cases, a linking element could be incorporated before the post-modifier, e.g., *chemin-de-fer* (French, lit. '*road of iron*', i.e., '*railway*').

Romanian, the only contemporary Romance language with grammatical cases, also allows for noun compounds formed using genitive, which is assigned by the definite article of the first noun to the second one, and is realized as a suffix if the second noun is preceded by the definite article, e.g., *frumuseţea fete*<u>*i*</u> (lit. '*beauty-the girl-gen*', meaning '*the beauty of the girl*'). (Romalo & *al.* 2005; Girju 2006; Girju 2007a; Girju 2007b)

2.3.3 Slavic Languages

Like in English, the NP structure of the Slavic languages is predominantly right-headed and adjectives precede the nouns they modify, e.g., *zelenaya trava* (Russian, '*green grass*'). Lexicalized noun compounds are written concatenated or with hyphens and are generally right-headed as well, e.g., *gorod-geroy* (Russian, '*city-hero*'). Nonlexicalized noun compounds, however, are left-headed, e.g., *uchebnik istorii* (Russian, lit. '*book history-gen*', i.e. '*history book*'). They are also case-inflected since Slavic languages are synthetic (except for the analytical Bulgarian and Macedonian[4]).

While here I use the term noun compound for both lexicalized and nonlexicalized noun compounds, traditionally, Russian grammar books use that term for neither of them; *complex words* and *word combinations* are used instead (Rozental' 1967; Rozental' 1970; Rozental' 1977; Gochev *et al.* 1987; Dzhambazov *et al.* 1994; Atanassova 2001).

Russian

Russian forms lexicalized right-headed compounds, which are typically included in a dictionary as separate lexical entries, and are considered *complex words.* They can be written as a single word, possibly with a connecting vowel, e.g., *kinoteatr* (*'cinema theater'*) and *krovoobrashchenie* (*'blood circulation'*), or as two words connected with a hyphen, e.g., *vagon-restoran*, (lit. *'car-restaurant'*, meaning *'dining car (in a train)'*). In the former two cases, inflections are applied to the second word only, but in the latter case there is a choice between inflecting both nouns (preferred in writing) or the second noun only (preferred in spoken language). Russian is also very productive in right-headed lexicalized noun compounds of type *portmanteaux*, where one or more of the nouns is shortened. For example, in *zamdekan* (*'deputy dean'*), the modifier *zamestnik* (*'deputy'*) is shortened to *zam.*

Much more interesting are the nonlexicalized noun compounds, considered by the grammars as *word combinations.* They are right-headed and the modifier is typically in *genitive* case[5], which can express a relation of possession of the head by the modifier,

[4]There is a heated linguistic (and political) debate about whether Macedonian is a separate language or is a regional literary form of Bulgarian. Since no clear criteria exist for distinguishing a dialect from a language, linguists remain divided on that issue.

[5]Such genitive structures are common in many other languages, including German, Greek, and Latin.

e.g., *kniga uchenika* is formed out of *kniga* + *uchenik* + *a*-gen (lit. '*book student-gen*', i.e. '*student's book*'). Genitive is also used in part-whole relations (e.g., *krysha doma*, lit. '*roof house-gen*', i.e. '*house's roof*'), in subject nominalizations (e.g., *lay sobaki*, lit. '*barking dog-gen*', i.e. '*dog's barking*'), in object nominalizations (e.g., *reshenie zadachi*, lit. '*solution problem-gen*', i.e. '*problem's solution*'), for measurements (e.g., *butylka vodki*, lit. '*bottle vodka-gen*', i.e. '*bottle of vodka*'), and for sets (e.g., *gruppa detey*, lit. '*group children-gen*', i.e. '*group of children*'),

While genitive is the most frequent case for the modifier, other cases are also possible, depending on the grammatical function and the semantic relation between the head and the modifier.

For example, *dative* case is used when the head is an object and the modifier is the recipient of that object, e.g., *pamyatnik partizanam*, (lit. '*monument partisans-dat*', i.e. '*monument to the partisans*'). It is also used in slogans, e.g., *slava geroyam*, (lit. '*glory heroes-dat*', i.e. '*glory to the heroes*').

Instrumental case is possible as well: to do an comparison (e.g., *nos kartoshkoy*, lit. '*nose potato-instr*', i.e. '*nose that looks like a potato*'), to express an instrument applied to the head (e.g., *porez nozhom*, lit. '*cut knife-instr*', i.e. '*cut with a knife*'), to express a mean of transportation (e.g., *poezdka avtobusom*, lit. '*travel bus-instr*', i.e. '*travel with a bus*'), to describe circumstances (e.g., *ezda polem*, lit. '*ride field-instr*', i.e. '*ride in the field*'), in subject nominalizations (e.g., *proverka uchitelem*, lit. '*check teacher-instr*', i.e. '*checking by the teacher*'), and in object nominalizations (e.g., *zanyatie baletom*, lit. '*occupation ballet-instr*', i.e. '*occupation with ballet*'). In the latter case, instrumental focuses on the action, while genitive focuses on the theme; compare *komandovanie polkom* (lit. '*commanding regiment-instr*', i.e. '*the action of commanding the regiment*') and *komandovanie polka* (lit. '*command regiment-gen*', i.e. '*the people who command the regiment*').

In all above examples, the head noun is in nominative. If a noun compound has to be inflected by case, the corresponding inflectional changes are applied to the head only (the first noun), e.g., *kniga uchenika* in accusative becomes *knigu uchenika*.

As in English, noun compounds can be much longer than two words. For example, *obsuzhdenie kriteriev ocenki kachestva obucheniya* (lit. '*discussion criteria-gen evaluation-gen quality-gen education-gen*', meaning '*discussion on the criteria for the evaluation of the quality of education*'), where all modifiers are in genitive. From a syntactic viewpoint, this is a right-bracketed noun compound:

[*obsuzhdenie* [*kriteriev* [*ocenki* [*kachestva obucheniya*]]]]

In Russian, different case patterns correspond to different syntactic structures:

- **head+gen+gen** is generally right-bracketed (as we saw above).

- **head+instr+gen** is right-bracketed. For example, *voshishchenie zelen'yu poley*, lit. '*admiration greenery-instr fields-gen*', meaning '*admiration for the greenery of the fields*'.

- **head+gen+instr** is left-bracketed. For example, *zahvat vlasti kontrrevolyucionerami*, lit. '*grab power-gen counter-revolutionary-instr*', meaning '*power grab by the counter-revolutionary*'.

Note that in some cases, the pattern **head+gen+gen** can be ambiguous. For example, *kniga istorii uchitelya* (lit. '*book (hi)story-gen teacher-gen*') can be interpreted as '*book about the teacher's story*' (right bracketing) or as '*history book of the teacher*' (left bracketing; preferred reading). As in English, in spoken language, stress and pause can help disambiguate between the two readings. However, *kniga uchitelya istorii* is not ambiguous; it has a right-bracketing interpretation only, meaning '*book of the history teacher*'.

Bulgarian

As in Russian, the highly lexicalized noun compounds in Bulgarian are right-headed and are written concatenated (often with a linking vowel), e.g., *kinosalon* (lit. '*cinema saloon*', meaning '*cinema hall*') and *gradonachalnik* (lit. '*city chief*', meaning '*mayor*'). (Boyadzhiev *et al.* 1998; Osenova & Simov 2007)

However, unlike Russian, which is synthetic, Bulgarian is an analytical language; therefore, head-modifier relations between nouns that are expressible with cases in Russian (i.e., following the pattern $N_{head} + N_{mod\_cased}$) would typically be expressed using prepositions in Bulgarian (i.e., following the pattern N_{head} + preposition + N_{mod}), which severely limits the number of nonlexicalized noun compounds in Bulgarian compared to Russian.

Nonlexicalized noun compounds in Bulgarian are pre-dominantly left-headed if written as two-words (often connected with a dash), e.g., *strana-chlenka* (lit. '*state-member*', meaning '*member-state*') and *ochi-chereshi* (lit. '*eyes-cherries*', meaning '*eyes that look like cherries*'), but there are also right-headed ones, e.g., *maystor gotvach* (lit. '*master cook*', meaning '*chef*') and *kandidat-student* (lit. '*candidate student*', meaning '*person who applies in an university to become a student*'). This distinction can be seen in definite forms, since the definite article in Bulgarian is a suffix and attaches to the head of the noun compound.

For example, in *stranata chlenka* and *ochite-chereshi* it is left-attached, but in *maystor gotvachyt* and *kandidat-studentyt* it is right-attached.

Interestingly, borrowed noun compounds keep the attachment from the language they are borrowed from. For example, *shkembe chorba* (‘*tripe soup*’), which comes from the Turkish *işkembe çorbası*, and the translations from English *ofis paket* (‘*office package*’), *biznes oferta* (‘*business offer*’), *Internet dostavchik* (‘*Internet provider*’), *onlayn usluga* (‘*on-line service*’), *general-mayor* (‘general-major’) and *ski avtomat* (‘*ski automaton*’) are all right-attached, which can be seen in the definite forms: *shkembe chorbata*, *ofis paketyt*, *biznes ofertata*, *Internet dostavchikyt*, *onlayn uslugata*, *general-mayoryt*, and *ski avotmatyt*.

2.3.4 Turkic Languages

In Turkish, the head in a noun compound is marked with a possessive suffix, e.g., *göbek dansı* (‘*belly dance*’) is formed out of *göbek* + *dans* + *ı*-poss. The possessive suffix is subject to vowel harmony constraints, and in some cases it can alter the last consonant of the word it attaches to, e.g., in *su bardağı* (‘*water glass*’), the original head word is *bardak*.

It is important to note that unlike genitive in Romanian or Russian, the possessive suffix in Turkish marks the head (the possessed), rather than the modifier (the possessor). Turkish also has a genitive case, which can be used to mark the possessor when necessary, just like English uses genitive in compounds like *Peter's book*. A corresponding example in Turkish would be *Ayşe'nin kitabı* (lit. ‘*Ayshe-gen book-poss*’, meaning ‘*Ayshe's book*’). As in English, the modifier does not have to be a person's name; it can be a regular noun as well, e.g., *manavın merakı*, which is formed out of *manav* (‘*greengrocer*’) + *merak* (‘*curiosity*’), meaning ‘*greengrocer's curiosity*’.

A noun compound like *göbek dansı* can in turn be used as a modifier of another noun, e.g., *kurs* (‘*course*’), which yields the left-bracketed double-marked noun compound *göbek dansı kursu* (‘*belly dance course*’). By contrast, a right-bracketed noun compound like *Berkeley mangal partisi* (‘*Berkeley BBQ party*’) has only one possessive suffix. Its derivation first puts together *mangal* and *parti* to form *mangal partisi*, which in turn is modified by *Berkeley*. Note that *parti* does not acquire a second suffix in the process since Turkish does not allow double possessive suffixes on the same word; in some cases, this can create ambiguity for longer compounds, e.g., *Berkeley göbek dansı kursu* is structurally ambiguous between bracketings (3) and (4):

(3) [[*Berkeley* [*göbek dansı*]] *kursu*]
(4) [*Berkeley* [[*göbek dansı*] *kursu*]]

2.4 Noun Compounds: My Definition, Scope Restrictions

Definition. *Following Downing (1977), I define a* noun compound *as a sequence of nouns which function as a single noun. I further require that all nouns in a noun compound be spelled as separate words. Occasionally, I will use the term* noun-noun compound *for noun compounds consisting of two nouns.*

Note that, according to my definition, not all sequences of nouns represent noun compounds; they have to function as a single noun. For example, *cat food* is not a noun compound in the sentence “*I gave my cat food.*”, which is perfectly reasonable: the two nouns are two separate arguments of the verb, a direct and an indirect object, and therefore they do not function as a single noun. Of course, *cat food* can be a noun compound in other contexts, e.g., in the sentence “*I would never allow my dog eat cat food.*”

The last requirement of my definition is computational rather than linguistic: I prefer to think of a noun compound as a sequence of words, each of which represents a noun, and I do not want to look inside the words. Therefore, *silkworm*, *textbook* and *headache* are single words for me, rather than noun compounds. However, they are noun compounds under most linguistic theories, and the fact that they are spelled as single words is rather a convention (but also a way to suggest a higher degree of lexicalization compared to noun compounds that are spelled separately). For example, *healthcare* and *health care* represent the same noun compound under most linguistic theories; the former probably expresses the writer's belief in a higher degree of lexicalization. However, under my definition, *healthcare* is not a noun compound at all while *health care* is, which creates some inconsistency, that I need to accept for the sake of simplicity of my experiments.

On the positive side, the requirement for separate spelling effectively eliminates many of the problematic classes of noun compounds that other researchers typically exclude explicitly (Levi 1978), e.g., *silverfish*, which is a metaphorical name, where the fish is not really *silver*; the synecdochic *birdbrain*, which actually refers to a quality of a person; and coordinate structures like *speaker-listener*. All these examples are spelled as single words and therefore do not represent noun compounds. Genitive constructions like *cat's food* are automatically excluded as well: while arguably being a separate token, the genitive marker is not a noun. Other commonly excluded noun compounds include names like *Union Square*, *Bush Street*, *Mexico City*, or *Stanford University*. While not formally excluded by my definition, they are not present in the datasets I use in my experiments.

2.5 Linguistic Theories

2.5.1 Levi's Recoverably Deletable Predicates

One of the most important theoretical linguistic theories of the syntax and semantics of noun compounds is that of Levi (1978). The theory targets the more general class of *complex nominals*, a concept grouping together[6] the partially overlapping classes of *nominal compounds*[7] (e.g., *peanut butter*, *mountain temperature*, *doghouse*, *silkworm*), *nominalizations* (e.g., *American attack*, *presidential refusal*, *dream analysis*) and *nonpredicate noun phrases*[8] (e.g., *electric shock*, *electrical engineering*, *musical critism*). The head of a complex nominal is always a noun, but the modifier is allowed to be either a noun or an adjective. In Levi's theory, the three subclasses share important syntactic and semantic properties. For example, the nominal compound *language difficulties* is synonymous with the nonpredicate NP *linguistic difficulties*: despite the surface morphological differences, they share the same semantic structure since the adjectival modifier of a complex nominal is derived from an underlying noun.

Levi focuses on the syntactic and semantic properties of the complex nominals and proposes detailed derivations within a theory of generative semantics. The derivations are based on two basic processes: *predicate deletion* and *predicate nominalization.* Given a two-argument predicate, *predicate deletion* gets rid of the predicate, retaining its arguments only (e.g., '*pie made of apples*' → *apple pie*), while *predicate nominalization* forms a complex

[6] Levi's theory is limited to *endocentric* complex nominals and excludes *exocentric* ones like metaphorical names (e.g., *silverfish*), synecdochic (e.g., *birdbrain*), coordinate structures (e.g., *speaker-listener*), and names (e.g., *San Francisco*).

[7] Levi's *nominal compounds* roughly correspond to my noun compounds, with some minor differences, e.g., she allows for word concatenations as in *silkworm.*

[8] Nonpredicate NPs contain a modifying adjective that cannot be used in predicate position with the same meaning. For example, the NP *a solar generator* cannot be paraphrased as '**a generator which is solar*'.

RDP	Example	Subj./Obj.	Traditional Name
$\texttt{CAUSE}_1$	*tear gas*	object	causative
$\texttt{CAUSE}_2$	*drug deaths*	subject	causative
$\texttt{HAVE}_1$	*apple cake*	object	possessive/dative
$\texttt{HAVE}_2$	*lemon peel*	subject	possessive/dative
$\texttt{MAKE}_1$	*silkworm*	object	productive; constitutive, compositional
$\texttt{MAKE}_2$	*snowball*	subject	productive; constitutive, compositional
`USE`	*steam iron*	object	instrumental
`BE`	*soldier ant*	object	essive/appositional
`IN`	*field mouse*	object	locative [spatial or temporal]
`FOR`	*horse doctor*	object	purposive/benefactive
`FROM`	*olive oil*	object	source/ablative
`ABOUT`	*price war*	object	topic

Table 2.1: **Levi's recoverably deletable predicates (RDPs).** The third column indicates whether the modifier was the subject or the object of the underlying relative clause.

nominal whose head is a nominalization of the underlying predicate and the modifier is either the subject or the object of that predicate (e.g., '*the President refused general MacArthur's request*' → *presidential refusal*).

Predicate Deletion

According to Levi, there exists a very limited number of *Recoverably Deletable Predicates* (*RDPs*) that can be deleted in the process of transformation of an underlying relative clause into a complex nominal: five verbs (`CAUSE`, `HAVE`, `MAKE`, `USE` and `BE`) and four prepositions (`IN`, `FOR`, `FROM` and `ABOUT`). See Table 2.1 for examples and alternative names for each predicate. While typically the modifier is derived from the object of the underlying relative clause, the first three verbs also allow for it to be derived from the subject.

Below I give an eight-step derivation of the complex nominal *musical clock*, using Levi's original notation. Step `b` forms a compound adjective, step `c` inserts a copula, step `d` forms a relative clause, step `e` deletes WH-*be*, step `f` performs a predicate preposing, step `g` deletes $\texttt{MAKE}_1$, and step `h` performs a morphological adjectivalization. This last step reveals the connection between two subclasses of complex nominals: nominal compounds and nonpredicate NPs.

```
a.  clock ## clock make music
b.  clock ## clock music-making
c.  clock ## clock be music-making
d.  clock ## which ## be music-making
e.  clock music-making
f.  music-making clock
g.  music clock
h.  musical clock
```

Note that the names of Levi's RDPs in Table 2.1 are capitalized, which is to stress that what is important is the semantic structure rather than the presence of the particular predicate. For example, `IN` refers to a generalized location which can be spatial or temporal, concrete or abstract, and the RDP deletion operation can recognize as instances of `IN` not only *in*, but also preposition like *on*, *at*, *near*, etc., (e.g., *terrestrial life* means '*life on earth*', *polar climate* is '*climate near the pole*', and *night flight* is '*flight at night*'), or any verb expressing that kind of locative relation (e.g., *desert rat* means '*rat inhabiting the desert*', *water lilies* means '*lilies growing in water*', etc.). Below I briefly describe each of the predicates:

$\texttt{CAUSE}_1$ means 'N_2 causes N_1'. It is derived by deletion of the present participle *causing* (from 'N_1-*causing* N_2'), e.g., *tear gas* is obtained from '*tear-causing gas*' (and earlier in the derivation, from '*gas that causes tears*').

`CAUSE`$_2$ means 'N_2 is caused by N_1'. It is derived by deletion of the past participle *caused* (from 'N_1-*caused* N_2'), e.g., *drug deaths* is obtained from '*drug-caused deaths*' (and earlier in the derivation, from '*deaths that are caused by drugs*').

`HAVE`$_1$ means 'N_2 has N_1'. For example, *apple cake* can be derived from the paraphrase '*cake with apples*'. Note however that the following intermediate derivation steps are not possible: *'*apple-having cake*' and *'*cake that is apple having*'. The verb can be replaced by a genitive marker, by prepositions like *of* and *with*, etc.

`HAVE`$_2$ means 'N_1 has N_2'. For example, *lemon peel* can be derived from the paraphrase '*peel of a lemon*' or '*lemon's peel*', but the following intermediate derivation steps are not possible: *'*lemon-had peel*' and *'*peel that is lemon-had*'. The verb can be replaced by a genitive marker, by prepositions like *of* and *with*, etc.

`FROM`. In this RDP, the modifier is a source for the head. Possible paraphrases include 'N_2 *from* N_1', 'N_2 *derived from* N_1', etc. The modifier is typically a natural object such as a vegetable (as in *cane sugar*) or an animal (as in *pork suet*), and the head denotes a product or a by-product obtained by processing the object named by the modifier. Another big subgroup is exemplified by *country visitor*, where the modifier denotes a previous location; it is ambiguous with a derivation from past tense + `IN` deletion. A third subgroup contains plants or animals separated by their biological source, like *peach pit*; it is ambiguous with a derivation by means of `HAVE`$_2$ deletion.

`MAKE`$_1$ means 'N_2 physically produces, causes to come into existence N_1'. It is derived by deletion of the present participle *making* (from 'N_1-*making* N_2'), e.g., *silkworm* is obtained from *silk-making worm.* Some forms analyzed as `MAKE`$_1$ deletion can have an alternative analysis as `FOR` deletion, e.g., *music box*. These different analyses correspond to a semantic difference.

$\texttt{MAKE}_\mathbf{2}$ means 'N_2 made up/out of N_1'. It is derived by deletion of the past participle *made* from 'N_1-*made* N_2'. There are three subtypes of this RDP: (a) the modifier is a unit and the head is a configuration, e.g., *root system*; (b) the modifier represents a material and the head represents a mass or an artefact, e.g., *chocolate bar*; (c) the head represents human collectives and the modifier specifies their membership, e.g., *worker teams*. All members of subtype (b) have an alternative analysis as `BE` deletion, e.g., *bar chocolate* can be analyzed as '*bar which is made of chocolate*' or as '*bar which is chocolate*'.

`BE`. This RDP covers multiple semantic classes, including: (a) compositional, e.g., *snowball*; (b) genus-species, e.g., *pine tree*; (c) metaphorical, e.g., *queen bee*; (d) coordinate, e.g., *secretary-treasure*; and (e) reduplicated, e.g., *house house*. The last two classes are exocentric and thus excluded from Levi's theory.

`IN`. Possible paraphrases for this RDP include 'N_2 *be located at* N_1', 'N_2 *in* N_1', 'N_2 *on* N_1', 'N_2 *at* N_1', 'N_2 *during* N_1', etc. It refers to a concrete (e.g., *desert mouse*), an abstract (e.g., *professional specialization*) or a temporal location (e.g., *night flight*).

`ABOUT`. This RDP can be paraphrased as 'N_2 *about* N_1', 'N_2 *concerned with* N_1', 'N_2 *concerning* N_1', 'N_2 *dealing with* N_1', 'N_2 *pertaining to* N_1', 'N_2 *on the subject of* N_1', 'N_2 *on* N_1', 'N_2 *over* N_1', 'N_2 *on the subject of* N_1', 'N_2 *whose subject is* N_1', etc. For example *tax law*, *sports magazine*, *border crisis*. This is a very homogeneous RDP.

`USE`. The verb *use* can represent two different lexical items: one with an instrumental and another one with an agentive meaning. Only the former can be used in a `USE` deletion, e.g., '*clock using electricity*' (instrumental) can give rise to the nominal compound *electrical clock*, but '*villagers using electricity*' (agentive) are not **electrical villagers*. If the head represents an activity then 'N_2 *by means of* N_1' is also possible, in addition to 'N_2 *using* N_1', e.g., *shock treatment* can be paraphrased as both '*treatment using shock*' and '*treatment by means of shock*'. If the head represents an object, e.g., *steam iron*, a longer-form paraphrase might be required, such as 'N_2 *using* N_1 *in order to function*' or 'N_2 *functioning by means of* N_1'.

`FOR`. This RDP expresses purpose. The preposition *for* can be `FOR` deleted in the case of '*spray for bugs*' (*bug spray*, i.e., that kills them), and '*spray for pets*' (*pet spray*, i.e., that helps them), but this is blocked when it means favor, as in '*women for peace*' (**peace women*), or in case of semantically empty object marker, as in '*appeal for money*' (**money appeal*). Possible paraphrases include 'N_2 *be for* N_1', 'N_2 *be intended for* N_1', 'N_2 *for* N_1', 'N_2 *be used for* N_1', 'N_2 *be for V-ing* N_1', etc. The last one is exemplified by *administrative office*, which means '*office for handling administration*'. The intervening verb V is often predictable from the meaning of the head, but not always. In some rare cases, the paraphrase could be 'N_2 *is good for* N_1', e.g., *beach weather*.

Predicate Nominalization

The second operation in Levi's theory that produces complex nominals is *predicate nominalization*. The resulting complex nominals have a nominalized verb as their head, and a modifier derived from either the subject or the object of the underlying predicate.

Multi-modifier nominalizations retaining both the subject and the object as modifiers are possible as well. Therefore, there are three types of nominalizations depending on

	Subjective	**Objective**	**Multi-modifier**
Act	*parental refusal*	*dream analysis*	*city land acquisition*
Product	*clerical errors*	*musical critique*	*student course ratings*
Agent	—	*city planner*	—
Patient	*student inventions*	—	—

Table 2.2: **Levi's nominalization types with examples.**

the modifier, which are combined with the following four types of nominalizations the head can represent: *act*, *product*, *agent* and *patient*. See Table 2.2 for examples.

Discussion

Levi's theory is one of the most sound and detailed theories on the syntax and semantics of noun compounds, but is not without drawbacks. First, it excludes many types of noun compounds (but it is normal for a theory to be limited in scope in order to stay focused). Second, despite the strong linguistic evidence for her decision to allow for adjectival modifiers in a noun compound, this remains controversial and is not widely accepted. Third, while being useful from a generative semantics point of view, her recoverably deletable predicates are quite abstract, which limits their value from a computational linguistics point of view, where '*lives in*' would arguably be more useful than `IN`. Finally, the dataset she built her theory on is quite heterogeneous and the noun compounds are of various degrees of lexicalization (compare *lemon peel* vs. *silkworm*); she also treated as two-word long some noun compounds which are actually composed of three or even four words, e.g., *wastebasket category*, *hairpin turn*, *headache pills*, *basketball season*, *testtube baby*, *beehive hairdo*.

Preposition	Example	Meaning
OF	*state laws*	laws of the state
FOR	*baby chair*	chair for babies
IN	*morning prayers*	prayers in the morning
AT	*airport food*	food at the airport
ON	*Sunday television*	television on Sunday
FROM	*reactor waste*	waste from a reactor
WITH	*gun men*	men with guns
ABOUT	*war story*	story about war

Table 2.3: **Lauer's prepositions with examples.**

2.5.2 Lauer's Prepositional Semantics

Lauer (1995) proposes that the semantics of a noun compound can be expressed by the following eight prepositions: *of*, *for*, *in*, *at*, *on*, *from*, *with* and *about*. See Table 2.3. While being simple, this semantics is problematic since the same preposition can indicate several different relations, and conversely, the same relation can be paraphrased by several different prepositions. For example, *in*, *on*, and *at*, all can refer to both location and time.

2.5.3 Descent of Hierarchy

Rosario *et al.* (2002) assume a head-modifier relationship between the nouns in a noun-noun compound and an argument structure for the head, similar to the argument structure of the verbs and related to the *qualia structure* in the theory of the generative lexicon of Pustejovsky (1995). Under this interpretation, the meaning of the head determines what can be done to it, what it is made of, what it is a part of, and so on. For example, for the word *knife*, the possible relations (with example noun-noun compounds) include the following:

- (Used-in): *kitchen knife*, *hunting knife*
- (Made-of): *steel knife*, *plastic knife*
- (Instrument-for): *carving knife*
- (Used-on): *meat knife*, *putty knife*
- (Used-by): *chef's knife*, *butcher's knife*

Some relations are specific for limited classes of nouns, while other are more general and apply to larger classes. Building on this idea, Rosario *et al.* (2002) propose a semi-supervised approach for characterizing the relation between the nouns in a bioscience noun-noun compound based on the semantic category in a lexical hierarchy each of the nouns belongs to. They extract all noun-noun compounds from a very large corpus (one million MEDLINE abstracts), and they make observations on which pairs of semantic categories the nouns tend to belong to. Based on these observations, they manually label the relations, thus avoiding the need to decide on a particular set of relations in advance.

They used the MeSH[9] (Medical Subject Heading) lexical hierarchy, where each concept is assigned a unique identifier (e.g., *Eye* is D005123) and one or more descriptor codes corresponding to particular positions in the hierarchy. For example, A (*Anatomy*), A01 (*Body Regions*), A01.456 (*Head*), A01.456.505 (*Face*), A01.456.505.420 (*Eye*). *Eye* is ambiguous and has a second code: A09.371 (A09 is *Sense Organs*).

The authors manually selected different levels of generalization, and then tested them on new data, reporting about 90% accuracy. For example, all noun-noun compounds

[9]http://www.nlm.nih.gov/mesh

in which the first noun is classified under the A01 sub-hierarchy (*Body Regions*), and the second one falls into A07 (*Cardiovascular System*), are hypothesized to express the same relation, e.g., *limb vein*, *scalp arteries*, *shoulder artery*, *forearm arteries*, *finger capillary*, *heel capillary*, *leg veins*, *eyelid capillary*, *ankle artery*, *hand vein*, *forearm microcirculation*, *forearm veins*, *limb arteries*, *thigh vein*, *foot vein*, etc.

The authors empirically found that the majority of the observed noun-noun compounds fall within a limited number of semantic category pairs corresponding to the top levels in the lexical hierarchy, e.g., A01-A07; most of the remaining ones require descending one or two levels down the hierarchy for at least one of the nouns in order to arrive at the appropriate level of generalization of the relation. For example, the relation is not homogeneous when the modifier falls under A01 (*Body Regions*) and the head is under M01 (*Persons*), e.g., *abdomen patients*, *arm amputees*, *chest physicians*, *eye patients*, *skin donor*; it depends on whether the person is a patient, a physician, or a donor. Making this distinction requires descending one level down the hierarchy for the head:

A01-M01.643 (*Patients*): *abdomen patients*, *ankle inpatient*, *eye outpatient*

A01-M01.526 (*Occupational Groups*): *chest physician*, *eye nurse*, *eye physician*

A01-M01.898 (*Donors*): *eye donor*, *skin donor*

A01-M01.150 (*Disabled Persons*): *arm amputees*, *knee amputees*

The idea of the descent of hierarchy is appealing and the demonstrated accuracy is very high, but it is not without limitations. First, the classification is not automated; it is performed manually. Second, the coverage is limited by the lexical hierarchy, most likely to specific domains. Third, there are problems caused by lexical and relational ambiguities.

Finally, the approach does not propose explicit names for the assigned relations.

The descent of hierarchy is similar in spirit to the work of Li & Abe (1998), who apply the *Minimum Description Length* (*MDL*) principle to the task of acquiring case frame patterns for verbs. The principle is a theoretically sound approach to model selection, introduced by Rissanen (1978). Its main idea is that learning represents a form of data compression: the better a model captures the principles underlying a given data sample, the more efficiently it can describe it. While a more complex model has a better chance of fitting the data well, being too specific it could also overfit it and therefore miss some important generalizations. The MDL principle states that the optimal balance between the complexity of the model and the fit to the data is achieved by minimizing the sum of the number of bits needed to encode the model (*model length*) and the number of bits needed to encode the data under that model (*data description length*). See (Quinlan & Rivest 1989) for details. Li & Abe (1998) apply the MDL principle to acquiring case frame patterns for verbs. Given a verb, e.g., *fly* and a set of example arguments with corresponding corpus frequencies, e.g., *crow*:2, *eagle*:2, *bird*:4, *bee*:2 for one of its slots, e.g., the direct object, they try to find the best level of generalization over the possible values of that slot in terms of *WordNet* categories. For example, in the case of the verb *fly*, suitable categories include `BIRD` and `INSECT`, but not the more general `ANIMAL`. Abe & Li (1996) and Li & Abe (1999) learn dependencies between case frame slots in a similar manner. Wagner (2005) extends their work and applies it to the task of learning thematic role relations at the appropriate level of generalization, e.g., `FOOD` is a suitable *patient* for the verb *eat*, while `CAKE` and `PHYSICAL_OBJECT` are not.

Chapter 3

Parsing Noun Compounds

In this chapter, I describe a novel, highly accurate lightly supervised method for making left vs. right bracketing decisions for three-word noun compounds. For example, [[*tumor suppressor*] *protein*] is left-bracketed, while [*world* [*food production*]] is right-bracketed. Traditionally, the problem has been addressed using unigram and bigram frequency estimates used to compute adjacency- and dependency-based word association scores. I extend this approach by introducing novel surface features and paraphrases extracted from the Web. Using combinations of these features, I demonstrate state-of-the-art results on two separate collections, one consisting of terms drawn from encyclopedia text, and another one extracted from bioscience text.

These experiments were reported in abbreviated form in (Nakov & Hearst 2005a) and (Nakov *et al.* 2005a).

3.1 Introduction

The semantic interpretation of noun compounds of length three or more requires that their syntactic structure be determined first. Consider for example the following contrastive pair of noun compounds:

(1) *liver cell antibody*
(2) *liver cell line*

In example (1), there is an *antibody* that targets a *liver cell*, while example (2) refers to a *cell line* which is derived from the *liver*. In order to make these semantic distinctions accurately, it can be useful to begin with the correct grouping of terms, since choosing a particular syntactic structure limits the options left for semantics. Although equivalent at the part of speech (POS) level, the above two noun compounds have different constituency trees, as Figure 3.1 shows.

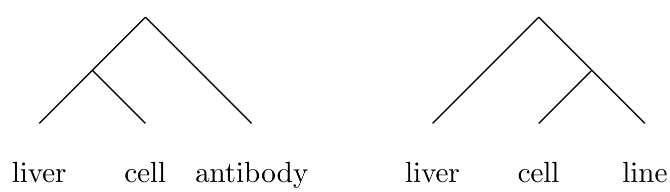

Figure 3.1: **Left vs. right bracketing:** constituency trees.

The above trees can be represented using brackets, which gives the name of the problem, *noun compound bracketing*:

(1b) [[*liver cell*] *antibody*] (left bracketing)
(2b) [*liver* [*cell line*]] (right bracketing)

Longer noun compounds like *colon cancer tumor suppressor protein* are rarely dealt with, since it is believed that parsing them can be reduced to similar left/right-bracketing decisions for triples of nouns. For example, suppose we have decided that [*colon cancer*] and [*tumor suppressor*] are noun compounds and are used as subunits in the bracketing: [*colon cancer*] [*tumor suppressor*] *protein*. Assuming a noun compound behaves like its head, we end up with a bracketing problem for the compound *cancer suppressor protein*. If we decide on a right bracketing for that compound, we end up with the following overall structure: [[*colon cancer*] [[*tumor suppressor*] *protein*]].

Parsing the noun compound is a necessary step towards semantic interpretation since the syntactic structure reveals the sub-parts between which relations need to be assigned, e.g., for the above example, we have the following semantic representation:

[*tumor suppressor protein*] which is implicated in [*colon cancer*] (`IN`; `LOCATION`)
[*protein*] that acts as [*tumor suppressor*] (`IS`; `AGENT`)
[*suppressor*] that inhibits [*tumor(s)*] (`OF`; `PURPOSE`)
[*cancer*] that occurs in [*(the) colon*] (`OF`; `IN`; `LOCATION`)

Below I describe a method for parsing three-word noun compounds in written text. In spoken text, in most situations there will be differences in stress or pauses that would help speakers to pick the correct interpretation (as I already mentioned in section 2.1). I further limit the present discussion to English; other languages, like Russian and Turkish, would use explicit case markers to help noun compound parsing (see section 2.3 for details). Finally, while I focus on noun compounds, similar ambiguities occur for adjectival modifiers, e.g., *graduate student parent*, *real world application*, *ancient Greek corpus*, *American Heart Association*.

3.2 Related Work

3.2.1 Unsupervised Models

The best known early work on automated unsupervised noun-compound bracketing is that of Lauer (1995), who introduces the probabilistic *dependency model* and argues against the *adjacency model*[1], used by Marcus (1980), Pustejovsky *et al.* (1993) and Resnik (1993). Given a three-word noun compound $w_1w_2w_3$, the adjacency model compares the strengths of association $Assoc(w_1, w_2)$ and $Assoc(\underline{w_2}, w_3)$, while the dependency model compares $Assoc(w_1, w_2)$ and $Assoc(\underline{w_1}, w_3)$.

Lauer collects n-gram statistics from *Grolier's encyclopedia*[2], which contain about eight million words (in 1995). To overcome data sparsity issues, he estimates probabilities over conceptual categories t_i in *Roget's thesaurus*[3] rather than for individual words w_i. Lauer defines the *adjacency* and the *dependency* models, respectively, as the following ratios:

$$R_{adj} = \frac{\sum_{t_i \in cats(w_i)} \Pr(t_1 \to t_2 | t_2)}{\sum_{t_i \in cats(w_i)} \Pr(t_2 \to t_3 | t_3)} \tag{3.1}$$

$$R_{dep} = \frac{\sum_{t_i \in cats(w_i)} \Pr(t_1 \to t_2 | t_2) \Pr(t_2 \to t_3 | t_3)}{\sum_{t_i \in cats(w_i)} \Pr(t_1 \to t_3 | t_3) \Pr(t_2 \to t_3 | t_3)} \tag{3.2}$$

where $cats(w_i)$ denotes the set of conceptual categories from *Grolier's encyclopedia* the word w_i can belong to, and $\Pr(t_i \to t_j | t_j)$ is the probability that the term t_i modifies a given term t_j, and is estimated as follows:

$$\Pr(t_i \to t_j | t_j) = \frac{\#(t_i, t_j)}{\#(t_j)} \tag{3.3}$$

---

[1] See sections 3.3.1, 3.3.2 and 3.3.3 for more details and a comparison between the two models.
[2] `http://go.grolier.com`
[3] `http://thesaurus.reference.com/`

Model	**AltaVista**	**BNC**
Baseline	63.93	63.93
$\#(n_1, n_2) : \#(n_2, n_3)$	77.86	66.39
$\#(n_1, n_2) : \#(n_1, n_3)$	**78.68**	65.57
$\#(n_1, n_2)/\#(n_1) : \#(n_2, n_3)/\#(n_2)$	68.85	65.57
$\#(n_1, n_2)/\#(n_2) : \#(n_2, n_3)/\#(n_3)$	70.49	63.11
$\#(n_1, n_2)/\#(n_2) : \#(n_1, n_3)/\#(n_3)$	80.32	66.39
$\#(n_1, n_2) : \#(n_2, n_3)$ (NEAR)	68.03	63.11
$\#(n_1, n_2) : \#(n_1, n_3)$ (NEAR)	71.31	67.21
$\#(n_1, n_2)/\#(n_1) : \#(n_2, n_3)/\#(n_2)$ (NEAR)	61.47	62.29
$\#(n_1, n_2)/\#(n_2) : \#(n_2, n_3)/\#(n_3)$ (NEAR)	65.57	57.37
$\#(n_1, n_2)/\#(n_2) : \#(n_1, n_3)/\#(n_3)$ (NEAR)	75.40	**68.03**

Table 3.1: **Noun compound bracketing experiments of Lapata & Keller (2004):** Accuracy in %s for *AltaVista* vs. *BNC* counts. Here $\#(n_i)$ and $\#(n_i, n_j)$ are estimates for word and bigram frequencies, respectively, and (NEAR) means that the words co-occur in a ten-word window. Shown in bold are the best performing models on the development set.

where $\#(t_i, t_j)$ and $\#(t_j)$ are the corresponding bigram and unigram frequencies calculated for the text of *Grolier's encyclopedia*.

These models predict a left bracketing if the ratio (R_{adj} or R_{dep}) is greater than one, and right bracketing if it is less than one. Lauer evaluates them on a testing dataset of 244 unambiguous noun compounds derived from *Grolier's encyclopedia* (see section 3.8.1 for a detailed description of the dataset and Table 3.3 for some examples). Below I will refer to that dataset as *Lauer's dataset*. The baseline of always predicting a left bracketing yields 66.8% accuracy, and the adjacency model is not statistically significantly different, obtaining only 68.9% accuracy. The dependency model achieves significantly better results, at 77.5% accuracy. Adding part-of-speech information in a more sophisticated *tuned* model allowed Lauer to improve the results to 80.70%.

More recently, Lapata & Keller (2004) and Lapata & Keller (2005) improve on Lauer's dependency results utilizing very simple word and bigram counts estimated using exact phrase queries against *AltaVista* or the `NEAR` operator[4], with all possible word inflections. They try several different models, and therefore reserve half of *Lauer's dataset* for model selection (tuning set), and test on the remaining 122 examples (testing set), which changes the accuracy of their left-bracketing baseline to 63.93%. Table 3.1 shows the models they experiment with and the corresponding results. The model that performs best on the tuning set is inspired from Lauer's dependency model and compares the bigram frequencies for $\#(n_1, n_2)$ and for $\#(n_1, n_3)$. On the test set, this model achieves 78.68% accuracy, which is a small improvement compared to the 77.50% for the dependency model in Lauer's experiments, but is worse compared to Lauer's *tuned* model. For comparison purposes, Lapata & Keller (2004) also try using frequency estimates from the *British National Corpus* (*BNC*), which represents 100M words (compared to 8 million in Lauer's experiments) of carefully edited, balanced English text. As Table 3.1 shows, the results when using *BNC* are much worse compared to using *AltaVista*: the best model that uses *BNC* only achieves 68.03% accuracy, which is statistically worse compared to the best model using *AltaVista* (78.68%). This result confirms the observation of Banko & Brill (2001) that using orders of magnitude more data can lead to significant improvements: the Web is orders of magnitude bigger than *BNC*.

3.2.2 Supervised Models

Girju *et al.* (2005) propose a *supervised* model for noun compound bracketing, based on 15 semantic features: five distinct semantic features calculated for each of the three nouns. The features require the correct *WordNet* sense for each noun to be provided:

[4] In *AltaVista*, a query for `x NEAR y` forces `x` to occur within ten word before or after `y`.

1. ***WordNet* derivationally related form**. Specifies if that sense of the noun is related to a verb in *WordNet*. For example, in "*coffee maker industry*", the correct sense of the second noun is *maker#3*, which is related to the verb to *make#6*.

2. ***WordNet* top semantic class of the noun**. For example, in "*coffee maker industry*", *maker#3* is a {*group, grouping*}*#1*.

3. ***WordNet* second top semantic class of the noun**. For example, in "*coffee maker industry*", *maker#3* is a *social_group#1*.

4. ***WordNet* third top semantic class of the noun.** For example, in"*coffee maker industry*", *maker#3* is *organizational#1*.

5. **Nominalization.** Indicates if the noun is a nominalization[5]. For example, in "*coffee maker industry*", *maker* is a nominalization.

Since their model is supervised, Girju *et al.* (2005) need training data. Therefore they assemble 49,208 sentences from *Wall Street Journal* articles from the Question Answering track of TREC-9[7] in 2000. Then they use Lauer's (1995) heuristic to extract candidate three-word noun compounds from that text, looking for sequences of three nouns not preceded and not followed by other nouns. The extracted candidates are checked for errors and manually bracketed in context by two Ph.D. students in Computational Semantics. This procedure yields 362 examples after agreement between the annotators has been reached.

Girju *et al.* (2005) use these 362 examples as training data, and the original *Lauer's*

[5]A noun is considered a nominalization if it is listed in the *NomLex*[6] dictionary of nominalizations (Macleod *et al.* 1998), or if it is an *event* or an *action* in *WordNet*.

[7]TExt Retrieval Conference, (`http://www.trec.nist.gov/`).

dataset of 244 examples as test data. Due to major differences between the training and the test datasets, they only achieve 73.10% accuracy, which is worse than the best unsupervised results of Lauer (1995) and Lapata & Keller (2004). Therefore, Girju *et al.* (2005) mix the training and the testing set and then randomly create new *shuffled* training and test sets of the same sizes as before: 362 training and 244 testing examples. Using these *shuffled* datasets, they achieve 83.10% accuracy with the C5.0 decision tree classifier (Quinlan 1993), which represents an improvement over Lauer (1995) and Lapata & Keller (2004).

For comparison purposes, Girju *et al.* (2005) repeat the Web-based experiments of Lapata & Keller (2004) on the *shuffled* test dataset (using *Google*), which yields 77.36% accuracy for the *dependency* model, and 73.45% for the *adjacency* model. Table 3.2 allows for an easy interpretation of these supervised experiments in comparison with the unsupervised experiments of Lauer (1995) and Lapata & Keller (2004).

Since having the correct *WordNet* sense might be unrealistic, Girju *et al.* (2005) also try using the first *WordNet* sense instead of the manually assigned one. As Table 3.2 shows, this causes a significant drop in accuracy on the *shuffled* dataset – from 83.10% to 74.40% – which suggests the algorithm may be sensitive to whether the provided *WordNet* sense is correct or not.

3.2.3 Web as a Corpus

In 2001, Banko & Brill (2001) advocated the creative use of very large text collections as an alternative to sophisticated algorithms and hand-built resources. They demonstrated the idea on a lexical disambiguation problem for which labeled examples were available "for free". The problem was to choose which of two to three commonly confused words,

Publication	**Training**	**Testing**	**Model**	**Acc.**
(Lauer 1995)		Lauer: 244	baseline (*left*)	66.80
			adjacency	68.90
			dependency	**77.50**
			tuned	**80.70**
(Lapata & Keller 2004)		Lauer: 122	baseline (*left*)	63.93
			best *BNC*	68.03
			best *AltaVista*	**78.68**
(Girju *et al.* 2005)	additional: 362	Lauer: 244	baseline (*left*)	66.80
			C5.0	***73.10**
			C5.0 (no WSD)	*72.80
(Girju *et al.* 2005)	shuffled: 362	shuffled: 244	baseline (*left*)	66.80
			C5.0	***83.10**
			C5.0 (no WSD)	*74.40
			Google adjacency	73.45
			Google dependency	77.36

Table 3.2: **Noun compound bracketing experiments on the *Lauer's dataset*:** accuracy in %s. Lauer (1995) uses no training nor tuning data, and tests on all 244 examples. Lapata & Keller (2004) use 122 of the examples for tuning and the remaining 122 ones for testing. Girju *et al.* (2005) present supervised models which are trained on 362 additional examples; the shuffled datasets are mixes of *Lauer's dataset* and these additional examples. The results for the supervised models are marked with an asterisk.

e.g., {*principle, principal*}, are appropriate for a given context. The labeled data was "free" since the authors could safely assume that, in the carefully edited text in their training set, the words are usually used correctly. They demonstrated that even using a very simple algorithm, the results continued to improve log-linearly with more training data, even out to a billion words. They also found that even the worst algorithm, with little training data, performed well when given orders of magnitude more data. Therefore, they concluded that obtaining more training data may be more effective overall than devising more sophisticated algorithms.

The question then arises about how and whether to apply this idea more generally for a wide range of natural language processing tasks. Today, the obvious answer is to use the Web.

Using the Web as a training and testing corpus is attracting ever-increasing attention. In 2003, the journal *Computational Linguistics* had a special issue on the Web as a corpus (Kilgariff & Grefenstette 2003). In 2005, the Corpus Linguistics conference hosted the first workshop on the *Web as Corpus* (*WAC*). In 2006, there was a WAC workshop in conjunction with the 11^{th} Conference of the European Chapter of the Association for Computational Linguistics, and in 2007 there was a WAC workshop, incorporating CleanEval.

The Web has been used as a corpus for a variety of NLP tasks including, but not limited to: machine translation (Grefenstette 1999; Resnik 1999a; Nagata *et al.* 2001; Cao & Li 2002; Way & Gough 2003), anaphora resolution (Modjeska *et al.* 2003), prepositional phrase attachment (Olteanu & Moldovan 2005; Volk 2001; Calvo & Gelbukh 2003), question answering (Soricut & Brill 2006; Dumais *et al.* 2002), extraction of semantic relations (Chklovski & Pantel 2004; Shinzato & Torisawa 2004; Idan Szpektor & Coppola 2004), language modeling (Zhu & Rosenfeld 2001; Keller & Lapata 2003), word sense disambiguation (Mihalcea & Moldovan 1999; Santamaría *et al.* 2003; Zahariev 2004).

Despite the variability of applications, the most popular use of the Web as a corpus is as a source of page hit counts, which are used as an estimate for n-gram word frequencies. Keller & Lapata (2003) have demonstrated a high correlation between page hits and corpus bigram frequencies as well as between page hits and plausibility judgments.

In a related strand of work, Lapata & Keller (2005) show that computing n-gram statistics over very large corpora yields results that are competitive with if not better than the best supervised and knowledge-based approaches on a wide range of NLP tasks. For example, they showed that for the problems of machine translation candidate selection, noun compound interpretation and article generation, the performance of an n-gram based model computed using search engine statistics was significantly better than the best supervised algorithm. For other problems, like noun compound bracketing and adjective ordering, they achieve results that are not significantly different from the best supervised algorithms, while for spelling correction, countability detection and prepositional phrase attachment they could not match the performance of the supervised state-of-the-art models. Therefore, they concluded that Web-based n-gram statistics should be used as the baseline to beat.

I feel the potential of these ideas is not yet fully realized, and I am interested in finding ways to further exploit the availability of enormous Web corpora as implicit training data. This is especially important for structural ambiguity problems in which the decisions must be made on the basis of the behavior of individual lexical items. The problem is to figure out how to use information that is latent in the Web as a corpus, and Web search engines as query interfaces to that corpus.

3.3 Adjacency and Dependency

3.3.1 Adjacency Model

According to the bracketing representation introduced above in section 3.1, given a three-word noun compound $w_1 w_2 w_3$, the task is to decide whether w_2 is more closely associated with w_1 or with w_3. Therefore, we need to compare the strength of association between the first two and the last two words, which is the **adjacency model** (Lauer 1995):

- if $Assoc(w_1, w_2) < Assoc(w_2, w_3)$, predict *right bracketing*;
- if $Assoc(w_1, w_2) = Assoc(w_2, w_3)$, make no prediction;
- if $Assoc(w_1, w_2) > Assoc(w_2, w_3)$, predict *left bracketing*.

3.3.2 Dependency Model

Lauer (1995) proposes an alternative *dependency model*, where there is an arc pointing from the modifier to the head it depends on, as shown in Figure 3.2.

Figure 3.2: **Left vs. right bracketing:** dependency structures.

In this representation, both the left and the right dependency structures contain a link $w_2 \to w_3$, but differ because of $w_1 \to w_2$ and $w_1 \to w_3$, respectively.[8] Therefore, the **dependency model** focuses not on w_1, rather than on w_2:

- if $Assoc(w_1, w_2) < Assoc(\underline{w_1}, w_3)$, predict *right bracketing*;
- if $Assoc(w_1, w_2) = Assoc(\underline{w_1}, w_3)$, make no prediction;
- if $Assoc(w_1, w_2) > Assoc(\underline{w_1}, w_3)$, predict *left bracketing.*

3.3.3 Adjacency vs. Dependency

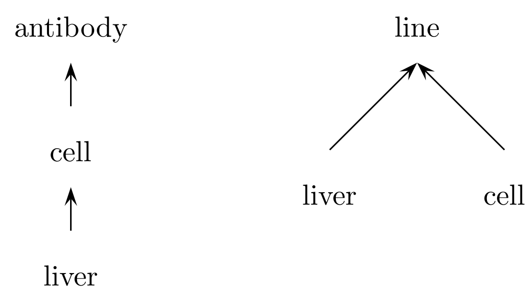

Figure 3.3: **Left vs. right bracketing:** dependency trees.

Figure 3.3 shows the *dependency trees* corresponding to the dependency structures from Figure 3.2. Note the structural difference from the constituency trees in Figure 3.1. First, the constituency trees contain words in the leaves only, while the dependency trees have words in the internal nodes as well. Second, the constituency trees are *ordered binary trees*: each internal node has exactly two *ordered* descendants, one left and one right, while there is no such ordering for the dependency trees.

[8]In my examples, the arcs always point to the right, i.e. the head always follows the modifier. While this is the typical case for English, it is not always the case, e.g., in *hepatitis b* the head is *hepatitis*. Since such cases are rare, below we will make the simplifying assumption that the modifier always precedes the head.

Consider examples (3) and (4): both are right-bracketed, but the order of the first two words is switched.

(3) [*adult* [*male rat*]] (right bracketing)
(4) [*male* [*adult rat*]] (right bracketing)

Despite (3) and (4) being different, the corresponding dependency structures are equivalent as Figures 3.4 and 3.5 show: there are no dependency arcs between *adult* and *male*, and therefore changing their linear order does not alter the dependency structure.

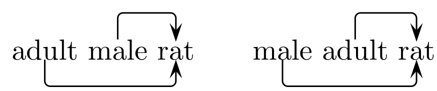

Figure 3.4: Dependency structures for *adult male rat* and for *male adult rat*.

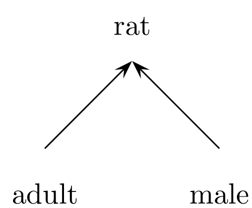

Figure 3.5: Shared dependency tree for *adult male rat* and for *male adult rat*.

The adjacency and the dependency models target different kinds of right-bracketed structures. Consider for example *home health care*: *health care* is a compound, which in turn is modified by *home* as a whole, which can easily be seen in the alternative spelling *home healthcare*, where the last two words are concatenated. This is different from *adult male rat*, where the two modifiers *adult* and *male* can be switched freely, which suggests that *adult* and *male* independently modify *rat*. We can conclude that, given a three-word noun compound $w_1 w_2 w_3$, there are two reasons it may take on right bracketing, $[w_1[w_2 w_3]]$:

(a) w_2w_3 is a compound, which is modified by w_1;

(b) w_1 and w_2 independently modify w_3.

Let us now look closely at the *adjacency* and the *dependency* models:

- **adjacency model:** compare $Assoc(w_1, w_2)$ and $Assoc(\underline{w_2}, w_3)$;

- **dependency model:** compare $Assoc(w_1, w_2)$ and $Assoc(\underline{w_1}, w_3)$.

Therefore the adjacency model checks (a) – whether w_2w_3 is a compound, while the dependency model checks (b) – whether w_1 modifies w_3.

Note that there is only a modificational choice in case of left bracketing. If w_1 modifies w_2, then w_1w_2 is a noun compound, which now acts as a single noun to modify w_3. Consider for example *law enforcement agent*: *law* is a modifier of *enforcement*, together they form a noun compound *law enforcement*, which in turn, as a whole, modifies *agent*.

3.4 Frequency-based Association Scores

Below I describe different ways to estimate the strength of association between a pair of words (w_i, w_j) – frequencies, probabilities, pointwise mutual information, and χ^2.

3.4.1 Frequency

The simplest association score I use is frequency-based:

$$Assoc(w_i, w_j) = \#(w_i, w_j) \tag{3.4}$$

where $\#(w_i, w_j)$ is a bigram frequency ($1 \leq i < j \leq 3$), which can be approximated as the number of page hits returned by a search engine for the exact phrase query "w_i w_j".

In the experiments below, I sum the frequencies for all inflected forms $infl(w_j)$ for the word w_j:

$$Assoc(w_i, w_j) = \sum_{t_j \in infl(w_j)} \#(w_i, t_j) \tag{3.5}$$

For example, the calculation of the strength of association between *stem* and *cell* would add up #(*stem*, *cell*) and #(*stem*, *cells*).

3.4.2 Conditional Probability

I also use conditional probabilities as an association score:

$$Assoc(w_i, w_j) = \Pr(w_i \rightarrow w_j | w_j) \tag{3.6}$$

where $\Pr(w_i \rightarrow w_j | w_j)$ is the probability that w_i modifies the head (w_j, $1 \leq i < j \leq 3$). This probability can be estimated as follows:

$$\Pr(w_i \rightarrow w_j | w_j) = \frac{\#(w_i, w_j)}{\#(w_j)} \tag{3.7}$$

where $\#(w_i, w_j)$ and $\#(w_j)$ are the corresponding bigram and unigram frequencies, which can be approximated as the number of page hits returned by a search engine, as demonstrated by Keller & Lapata (2003). See chapter 7 for a discussion on the stability of such estimates.

In my experiments, I sum the frequencies for all inflected forms of the word w_j:

$$Assoc(w_i, w_j) = \frac{\sum_{t_j \in infl(w_j)} \#(w_i, t_j)}{\sum_{t_j \in infl(w_j)} \#(t_j)} \tag{3.8}$$

3.4.3 Pointwise Mutual Information

Another association score I use is the *pointwise mutual information* (PMI):

$$Assoc(w_i, w_j) = \text{PMI}(w_i, w_j) \tag{3.9}$$

which is defined as follows:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{\Pr(w_i, w_j)}{\Pr(w_i)\Pr(w_j)} \tag{3.10}$$

Here $\Pr(w_i, w_j)$ is the probability of the sequence "$w_i\ w_j$", and $\Pr(w)$ is the probability of the word w on the Web. Let N be the total number of words/bigrams on the Web, which I estimate to be about 8 trillion: *Google* indexes about 8 billion pages and I hypothesize that each one contains about 1,000 words/bigrams on average. Then the probabilities can be estimated as follows:

$$\Pr(w_i, w_j) = \frac{\#(w_i, w_j)}{N} \tag{3.11}$$

$$\Pr(w_i) = \frac{\#(w_i)}{N} \tag{3.12}$$

And therefore for PMI we obtain:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{N \times \#(w_i, w_j)}{\#(w_i)\#(w_j)} \tag{3.13}$$

Again, in my experiments, I sum the frequencies for all inflected forms:

$$Assoc(w_i, w_j) = \log \frac{N \times \sum_{t_j \in infl(w_j)} \#(w_i, t_j)}{\sum_{t_i \in infl(w_i)} \#(t_i) \sum_{t_j \in infl(w_j)} \#(t_j)} \tag{3.14}$$

Using the definition of conditional probability, we can re-write eq. 3.10, as follows:

$$\mathrm{PMI}(w_i, w_j) = \log \frac{\Pr(w_i \rightarrow w_j | w_j)}{\Pr(w_i)} \tag{3.15}$$

The dependency model compares $\mathrm{PMI}(w_1, w_2)$ and $\mathrm{PMI}(w_1, w_3)$. We have:

$$\mathrm{PMI}(w_1, w_2) = \log \frac{\Pr(w_1 \rightarrow w_2 | w_2)}{\Pr(w_1)} = \log \Pr(w_1 \rightarrow w_2 | w_2) - \log \Pr(w_1) \tag{3.16}$$

$$\mathrm{PMI}(w_1, w_3) = \log \frac{\Pr(w_1 \rightarrow w_3 | w_3)}{\Pr(w_1)} = \log \Pr(w_1 \rightarrow w_3 | w_3) - \log \Pr(w_1) \tag{3.17}$$

Since $\log \Pr(w_1)$ is subtracted from both expressions, we can ignore it and directly compare $\log \Pr(w_1 \rightarrow w_2|w_2)$ to $\log \Pr(w_1 \rightarrow w_3|w_3)$. Since the logarithm is a monotonic function, we can simply compare $\Pr(w_1 \rightarrow w_2|w_2)$ to $\Pr(w_1 \rightarrow w_3|w_3)$, i.e. for the *dependency model*, comparing PMIs is equivalent to comparing conditional probabilities. This is not true for the *adjacency model* though.

3.4.4 Chi-square

The last association score is *Chi-square* (χ^2), which uses the following statistics:

$A = \#(w_i, w_j)$;

$B = \#(w_i, \overline{w_j})$, number of bigrams where w_i is followed by a word other than w_j;

$C = \#(\overline{w_i}, w_j)$, number of bigrams, ending in w_j, whose first word is other than w_i;

$D = \#(\overline{w_i}, \overline{w_j})$, number of bigrams with first word that is not w_i and second is not w_j.

These statistics are combined in the following formula:

$$\chi^2 = \frac{N(AD - BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{3.18}$$

Here $N = A + B + C + D$ is the total number of bigrams on the Web (estimated as 8 trillion), $B = \#(w_i) - \#(w_i, w_j)$ and $C = \#(w_j) - \#(w_i, w_j)$. While it is hard to estimate D directly, it can be calculated as $D = N - A - B - C$. Finally, as for the other association scores, I sum the unigram and bigram frequencies for the inflected word forms: $\#(w_i) = \sum_{t_i \in infl(w_i)} \#(t_i)$, and $\#(w_i, w_j) = \sum_{t_j \in infl(w_j)} \#(w_i, t_j)$.

3.5 Web-Derived Surface Features

Consciously or not, authors sometimes disambiguate the noun compounds they write by using surface-level markers to suggest the correct meaning. For example, in some cases an author could write *law enforcement officer* with a dash, as *law-enforcement officer*, thus suggesting a left bracketing interpretation. I have found that exploiting such markers, when they occur, can be very helpful for making bracketing predictions. The enormous size of the Web facilitates finding them frequently enough to be useful.

Unfortunately, Web search engines ignore punctuation characters, thus preventing querying directly for terms containing hyphens, brackets, apostrophes, etc. For example, querying for `"brain-stem cell"` would produce the same result as for `"brain stem cell"`, which makes the number of page hits unusable for counting the occurrences of hyphenation. Therefore, I look for the features indirectly by issuing an exact phrase query for the noun compound and then post-processing the results (up to 1,000 per query). I collect the returned text snippets (typically 1-2 sentences or pars of sentences) and then I search for the surface patterns using regular expressions over the text. Each match increases the score for left or right bracketing, depending on which one the pattern favors.

While some of the features may be more reliable than others, I do not try to weight them. Given a noun compound, I make a bracketing decision by comparing the total number of left-predicting surface feature instances (regardless of their type) to the total number of right-predicting feature instances. This appears as *Surface features (sum)* in Tables 3.4 and 3.6. The surface features used are described below.

3.5.1 Dash

One very productive feature is the *dash* (hyphen). Consider for example *cell cycle analysis*. If we can find a version of it in which a dash occurs between the first two words, *cell-cycle*, this suggests left bracketing for the full compound. Similarly, the dash in *donor T-cell* favors right bracketing. Dashes predicting right bracketing are less reliable though, as their scope is ambiguous. For example, in *fiber optics-system*, hyphen's scope is not *optics*, but *fiber optics*. Finally, multiple dashes are unusable, e.g., *t-cell-depletion*.

3.5.2 Genitive Marker

The genitive ending, or *genitive marker* is another useful feature. Finding the phrase *brain's stem cells* suggests right bracketing for *brain stem cells*, while finding *brain stem's cells* suggests left bracketing. Note that these features can also occur combined, as in *brain's stem-cells*, where we have both a *dash* and a *genitive marker*.

3.5.3 Internal Capitalization

Internal capitalization is another highly reliable feature. For example, the capitalized spellings *Plasmodium vivax Malaria* and *plasmodium vivax Malaria* both suggest left bracketing for the noun compound *plasmodium vivax malaria*, while *brain Stem cells* and *brain Stem Cells* both favor right bracketing for *brain stem cells*.

This feature is disabled on Roman digits and single-letter words in order to prevent problems with terms like *vitamin D deficiency*, which is right-bracketed and where the capitalization is just a convention.

3.5.4 Embedded Slash

I also make use of *embedded slashes.* For example, in *leukemia/lymphoma cell*, the first word is an alternative of the second one and therefore cannot modify it, which suggests right bracketing for *leukemia lymphoma cell.*

3.5.5 Parentheses

In some cases we can find instances of the noun compound in which one or more words are enclosed in *parentheses*, e.g., *growth factor (beta)* or *(growth factor) beta*, both of which suggest left bracketing for *growth factor beta*; or *(brain) stem cells* and *brain (stem cells)*, which favor right bracketing for *brain stem cells.*

3.5.6 External Dash

External dashes, i.e. dashes to words outside the target compound can be used as weak indicators. For example, <u>*mouse-*</u>*brain stem cells* favors right bracketing for *brain stem cells*, while *tumor necrosis factor*<u>*-alpha*</u> suggests left bracketing for *tumor necrosis factor.*

3.5.7 Other Punctuation

Even a comma, a period or a colon (or any special character) can act as an indicator. For example, "*health care, provider*" or "*lung cancer: patients*" are weak predictors of left bracketing, showing that the author chose to keep two of the words together, separating out the third one, while "*home. health care*" and "*adult, male rat*" suggest right bracketing.

3.6 Other Web-Derived Features

Some features can be obtained directly as the number of page hits for a suitably formulated query. As these counts are derived from the entire Web, not just from the top 1,000 summaries, they are of different magnitude and therefore should not be added to the frequencies of the above surface features; I use them in separate models.

3.6.1 Genitive Marker – Querying Directly

In some cases, it is possible to query for *genitive markers* directly: while search engines do not index apostrophes, since they index no punctuation, the *s* in a genitive marker is indexed. This makes it possible to query for "*brain stem's cell*" indirectly by querying for "*brain stem s cell*". An *s* in this context is very likely to have been part of a genitive marker, and if we assume so, then it is possible to make a bracketing decision by comparing the number of times an *s* appears following the second word vs. the first word. For example, a query for "*brain s stem cell*" returns 4 results, and for "*brain s stem cells*" returns 281 results, i.e. there are a total of 285 page hits supporting left bracketing for *brain stem cell*. On the other hand, the right-predicting queries "*brain stem s cell*" and "*brain stem s cells*" return 2 and 3 results, respectively. Thus there are only 5 right-predicting page hits vs. 285 left-predicting, and therefore the model makes a left bracketing prediction.

Note: I conducted the original experiments in 2005, when *Google* seemed to ignore punctuation. It appears that it can handle some of it now. For example, at the time of writing, "*brain stem's cell*" return 701 results, all having the genitive marker *'s*.

3.6.2 Abbreviation

Abbreviations are another very reliable feature. For example, *"tumor necrosis factor (NF)"* suggests a right bracketing, while *"tumor necrosis (TN) factor"* favors a left bracketing for *"tumor necrosis factor"*. Since search engines ignore parentheses and capitalization, a query for *"tumor necrosis factor (NF)"* is equivalent to a query for *"tumor necrosis factor nf"*. Again, I compare the number of page hits where the abbreviation follows the second word to the number of page hits where the abbreviation follows the first word (using inflections). In general, this model is very accurate, but errors may occur when the abbreviation is an existing word (e.g., *me*), a Roman digit (e.g., *IV*), a state name abbreviation (e.g., *CA*), etc.

3.6.3 Concatenation

Concatenation Adjacency Model

Some noun compounds are often written concatenated. I exploit this in adjacency and dependency models. Let $concat(w_i, w_j)$ be the concatenation of the words w_i and w_j. Given a three-word noun compound $w_1 w_2 w_3$, the *concatenation adjacency model* compares the number of page hits for $concat(w_1, w_2)$ and $concat(w_2, w_3)$, using all possible inflections of w_2 and w_3, respectively. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise.

Consider for example the noun compound *health care reform*. *Google* returns 98,600,000 page hits for *healthcare*, and 33,000 more for *healthcares*: a total of 98,633,000. On the other hand, there are only 471 page hits for *carereform*, and 27 more for *carereforms*:

a total of 498. Therefore, the model predicts a left bracketing.

Concatenation Dependency Model

Given a three-word noun compound $w_1 w_2 w_3$, the *concatenation dependency model* compares the number of page hits for *concat*(w_1, w_2) and *concat*(w_1, w_3), using all possible inflections of w_2 and w_3, respectively. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise.

As we have seen above, *Google* returns a total of 98,633,000 for *healthcare* and *healthcares*. On the other hand, there are only 27,800 page hits for *healthreform*, and 82 more for *healthreforms*: a total of 27,882. Therefore, the model predicts a left bracketing for the noun compound *health care reform*.

Concatenation Adjacency Triples

Given a three-word noun compound $w_1 w_2 w_3$, the *concatenation adjacency triples* model compares the number of page hits for the exact phrase *Google* queries: "*concat*(w_1, w_2) w_3" and "w_1 *concat*(w_2, w_3) using all possible inflections of w_2 and w_3, and for w_1 and w_3, respectively. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise.

At the time of writing, *Google* returns 668,000 page hits for *"healthcare reform"*, 77,700 page hits for *"healthcare reforms"*, and no page hits for *"healthcares reform"* or *"healthcares reforms"*: a total of 745,700. On the other hand, there are only 289 page hits for *"health carereform"*, 15 page hits for *"health carereforms"*, and no page hits for *"health carereforms"* or *"healths carereforms"*: a total of 304. This supports a left bracketing prediction for the noun compound *health care reform*.

3.6.4 Wildcard

I also make use of the *Google*'s support for "*", a wildcard substituting a single word.[9] The idea is to see how often two of the words follow each other and are separated from the third word by some other word(s). This implicitly tries to capture paraphrases involving the two sub-concepts making up the whole. I propose wildcards adjacency and dependency models, using up to three asterisks, possibly with a re-ordering.

Wildcard Adjacency Model

Given a three-word noun compound $w_1 w_2 w_3$, the *wildcards adjacency model* compares the number of page hits for "w_1 w_2 * w_3" and "w_1 * w_2 w_3", using all possible inflections of w_2 and w_3, and for w_1 and w_3, respectively. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise. I also use a version of the model with two and with three asterisks.

Consider the following example, where only one asterisk is used. At the time of writing, *Google* returns 556,000 page hits for *"health care * reform"*, 79,700 page hits for *"health care * reforms"*, 1 page hit for *"health cares * reform"*, and no page hits for *"health cares * reforms"*: a total of 635,701. On the other hand, there are only 255,000 page hits for *"health * care reform"*, 17,600 page hits for *"health * care reforms"*, 1 page hit for *"health * care reforms"*, and no page hits for *"healths * care reforms"*: a total of 272,601. This supports a left bracketing prediction for the noun compound *health care reform*.

[9] While it should substitute a single word, *Google*'s "*" operator often substitutes multiple words.

Wildcard Dependency Model

Given a three-word noun compound $w_1w_2w_3$, the *wildcards dependency model* compares the number of page hits for "w_1 w_2 * w_3" and "w_2 * w_1 w_3", using all possible inflections of w_2 and w_3. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise. I try using up to three asterisks.

Consider the following example, where only one asterisk is used. At the time of writing, *Google* returns 556,000 page hits for *"health care * reform"*, 79,700 page hits for *"health care * reforms"*, 1 page hit for *"health cares * reform"*, and no page hits for *"health cares * reforms"*: a total of 635,701. On the other hand, there are only 198,000 page hits for *"care * health reform"*, 12,700 page hits for *"care * health reforms"*, 122,000 page hit for *"cares * health reform"*, and no page hits for *"cares * health reforms"*: a total of 332,700. This supports a left bracketing prediction for the noun compound *health care reform*.

Wildcard Reversed Adjacency Model

Given a three-word noun compound $w_1w_2w_3$, the *wildcards reversed adjacency model* compares the number of page hits for "w_3 * w_1 w_2" and "w_2 w_3 * w_1", using inflections for w_2 and w_3, and for w_1 and w_3, respectively. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise.

Consider the following example, where only one asterisk is used. At the time of writing, *Google* returns 824,000 page hits for *"reform * health care"*, 3 page hits for *"reform * health cares"*, 119,000 page hits for *"reforms * health care"*, and 2 page hits for *"reforms * health cares"*: a total of 943,005. On the other hand, there are only 255,000 page hits for *"care reform * health"*, 1 page hit for *"care reform * healths"*, 13,900 page hit for *"care reforms * health"*, and no page hits for *"care reforms * healths"*: a total of 268,901. This supports a left bracketing prediction for the noun compound *health care reform*.

Wildcard Reversed Dependency Model

Given a three-word noun compound $w_1 w_2 w_3$, the *wildcards reversed dependency model* compares the number of page hits for "w_3 * w_1 w_2" and "w_1 w_3 * w_2", using all possible inflections of w_2 and w_3. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise. I try using up to three asterisks.

Consider the following example, where only one asterisk is used. At the time of writing, *Google* returns 824,000 page hits for *"reform * health care"*, 3 page hits for *"reform * health cares"*, 119,000 page hits for *"reforms * health care"*, and 2 page hits for *"reforms * health cares"*: a total of 943,005. On the other hand, there are only 273,000 page hits for *"health reform * care"*, 122,000 page hit for *"health reform * cares"*, 10,800 page hit for *"health reforms * care"*, and no page hits for *"health reforms * cares"*: a total of 405,800. This supports a left bracketing prediction for the noun compound *health care reform*.

3.6.5 Reordering

I also try a simple *reordering* without inserting asterisks. Given a three-word noun compound $w_1 w_2 w_3$, the *reordering model* compares the number of page hits for "w_3 w_1 w_2" and "w_2 w_3 w_1", using all possible inflections of w_2 and w_3, and for w_1 and w_3, respectively. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise. For example, at the time of writing, *Google* returns 137,000 page hits for *“reform health care”*, 1,010 page hits for *“reforms health care”*, and no page hits for *“reform health cares”* or “*reforms health cares*”: a total of 138,010. On the other hand, there are only 23,300 page hits for *“care reform health”*, 1,720 page hits for *“care reforms health”*, and no page hits for *“care reform healths”* or *“care reforms healths”*: a total of 25,020. This supports a left bracketing prediction for *health care reform*.

3.6.6 Internal Inflection Variability

Further, I try to use the *internal inflection variability*. Consider the three-word noun compound *“tyrosine kinase activation”*. Since it is left-bracketed, the first two nouns form a two-word noun compound (which in turn modifies the third noun). Therefore, we could expect to see this two-word noun compound inflected in the context of the three-word noun compound, i.e. *“tyrosine kinases activation”*. If it were right bracketed, we would expect a possible inflection on the first noun: *“tyrosines kinase activation”*. The model compares the number of page hits for the internal inflections of the second noun vs. the first noun. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise.

For example, the left-predicting patterns for *“tyrosine kinase activation”* are *“tyrosine kinases activation”* and *“tyrosine kinases activations”*, yielding 882 + 0 = 882 page hits. On the other hand, there are no page hits for *“tyrosines kinase activation”* nor for *“tyrosines kinase activations”*. This supports a left bracketing prediction for the noun compound *tyrosine kinase activation*.

In case of plural nouns, I convert them to singular. Consider for example the noun compound *math problems solutions*. The left-predicting patterns are "*math problem_ solutions*" and "*math problem_ solution_*", which yield 3,720 + 2,920 = 6,640 page hits. The right-predicting patterns are "*maths problems solutions*" and "*maths problems solution_*", which yield 2,990 + 0 = 2,990 page hits. This supports a left bracketing prediction.

3.6.7 Swapping the First Two Words

Recall the *adult male rat* example, discussed in section 3.3.3. This is a right-bracketed noun compound, where the first two words independently modify the third one, and therefore we can swap them. For example, at the time of writing, *Google* returns 78,200 page hits for *adult male rat* and 1,700 for *male adult rat*. I exploit this literally: testing whether it is feasible to swap the first two words, in which case I make a right bracketing prediction.

More formally, given a three-word noun compound $w_1 w_2 w_3$, the *swapping the first two words model* checks the number of page hits for "w_2 w_1 w_3", using all possible inflections for w_1, w_3, and w_3. It predicts a left bracketing if the former is greater, a right bracketing if the latter is greater, and makes no prediction otherwise.

3.7 Paraphrases

Many researchers point out that the semantics of the relation between the nouns in a noun compound can be made overt by a paraphrase (Warren 1978; Lauer 1995). For example, an author describing the concept of *brain stem cell* may choose to write it in a more expanded manner, paraphrasing it as *stem cells from the brain*. Since *stem* and *cells* are kept together in that paraphrase, this suggests a right bracketing. On the other hand, if the expansion groups the first two nouns together, as in *cells from the brain stem*, this would support a left bracketing.

Not all noun compounds can be readily expressed with a prepositional paraphrase (Levi 1978). For example, the noun compound *skyscraper office building* is best expanded as a copula paraphrase, e.g., *office building that is a skyscraper*, which predicts a right bracketing. Another option is to use verbal paraphrases like *pain associated with arthritis migraine*, which supports a left bracketing for *arthritis migraine pain*.

Some researchers use prepositional paraphrases as a proxy for determining the semantic relation that holds between the nouns in a *two-word* noun compound (Girju *et al.* 2005; Lapata & Keller 2004; Lauer 1995). Lauer (1995) takes the idea literally, formulating the semantic interpretation task as selecting one of the following prepositions[10]: *of*, *for*, *in*, *at*, *on*, *from*, *with* and *about*. For example, a *night flight* is paraphrased as a *flight at night*, while *war story* is paraphrased as a *story about war*.

In contrast, I use paraphrases in order to obtain evidence supporting left or right *syntactic* bracketing decision for *three-word* noun compounds. Rather than trying to decide on a single best paraphrase, I issue queries using several different paraphrase patterns. I find how often each one occurs on the Web, add separately the number of page hits predicting a left versus right bracketing and then compare the two sums. As before, the bigger sum wins, and no decision is made in case of equal sums. While some prepositions may be better predictors than other ones (see (Girju *et al.* 2005) for a frequency distribution), I have just one feature for all prepositions.

[10]He excludes *like*, which is mentioned by Warren (1978).

Unfortunately, search engines lack linguistic annotations and therefore they do not support typed queries like "*stem cells VERB PREP DET brain*", where the uppercase placeholders stand for a verb, a preposition and a determiner, respectively. This makes the general verbal paraphrases prohibitively expensive – a separate query is needed for every combination of a verb, a preposition and a determiner.

To overcome this problem, I use a fixed set of *verbal prepositions*, i.e. passive verbs accompanied by a preposition acting as a complex preposition, e.g., *used in* or *made of*; a full list is given below.

More formally, given a three-word noun compound $w_1 w_2 w_3$, I use the following generalized left-predicting exact phrase queries which keep w_1 and w_2 together. Note that w_2 and w_3 can be number-inflected for singular or plural, which is indicated with the $'$ sign in the patterns below:

1. w_3' `PREP DET?` w_1 w_2'
2. w_3' `COMPL BE DET?` w_1 w_2'
3. w_3' `COMPL BE PREP DET?` w_1 w_2'

Similarly, the right-predicting paraphrases are generated using corresponding generalized exact phrase queries that keep w_2 and w_3 together (w_1 and w_3 can be inflected):

1. w_2 w_3' `PREP DET?` w_1'
2. w_2 w_3' `COMPL BE DET?` w_1'
3. w_2 w_3' `COMPL BE PREP DET?` w_1'

The above queries are instantiated in all possible ways by enumerating all inflected forms for w_1, w_2 and w_3 (where indicated with the $'$ sign) and by substituting the placeholders `COMPL` (a complementizer), `BE` (a form of the verb to be), `DET` (a determiner) and `PREP` (a preposition) using the following values[11]:

- `PREP`:
  - **Nonverbal prepositions:** *about, across, after, against, all over, along, alongside, amid, amidst, among, around, as, as to, aside, at, before, behind, beside, besides, between, beyond, by, close to, concerning, considering, down, due to, during, except, except for, excluding, following, for, from, in, in addition to, in front of, including, inside, instead of, into, like, near, of, off, on, onto, other than, out, out of, outside, over, per, regarding, respecting, similar to, through, throughout, to, toward, towards, under, underneath, unlike, until, up, upon, versus, via, with, within, without*;
  - **Verbal prepositions:** *associated with, caused by, contained in, derived from, focusing on, found in, involved in, located at, located in, made of, performed by, preventing, related to, used by, used in, used for.*
- `DET` – an optional element, which can be any of the following:
  - **Determiners:** *a, an, the*;
  - **Quantifiers:** *all, each, every, some*;

[11] In the process of generation, I use *WordNet* to check the number agreement. While wrong agreement does not generally harm the accuracy since it generates phrases that simply return no results, I still try to eliminate it in order to reduce the total number of necessary queries.

  - **Possessive pronouns:** *his, her, their*;
  - **Demonstrative pronouns:** *this, these.*

- `COMPL`
  - **Complementizers:** *that, which, who.*

- `BE`
  - **Copula:** *are, is, was, were.*

Some example left-predicting pattern instantiations for the noun compound *bone marrow cells* follow: *cell from the bone marrow, cells in the bone marrow, cells derived from the bone marrow, cells that are from the bone marrows, cell which is in a bone marrow*, etc.

The reason to include so many prepositions comes from my concern that the coverage of Lauer's eight prepositions plus the copula paraphrases might be insufficient. While the extended set of prepositions did not improve the coverage in my experiments, it boosted the accuracy by more than 2%.

One possible explanation is that the specialized biomedical terms are paraphrased more often, probably in order to be explained to readers that are new to the domain, while Lauer's noun compounds are extracted from popular encyclopedia text and, therefore most of them are in everyday use.

3.8 Datasets

3.8.1 Lauer's Dataset

Lauer's dataset is considered the benchmark dataset for the noun compound bracketing problem. It is listed in the appendix of Lauer's thesis (Lauer 1995), and consists of 244 unambiguous three-word noun compounds extracted from *Grolier's encyclopedia*. In order to extract his noun compounds, Lauer POS-tagged the encyclopedia text and then extracted three-word noun compound candidates $w_1w_2w_3$ by heuristically looking for five-word sequences $w_0w_1w_2w_3w_4$, where w_1, w_2 and w_3 are nouns and w_0, w_4 are not nouns (Lauer 1995). This heuristic yielded 625 three-word noun compound candidate instances. Since Lauer's learning models rely on a thesaurus, he discarded any noun compound that contained a word that was not in *Roget's thesaurus*, which left him with 308 candidates.

Lauer manually investigated these 308 examples and tried to annotate each one with a left or a right bracketing using as context the whole encyclopedia article the noun compound occurred in. In the process, he discovered that 29 examples (i.e. 9%) represented extraction errors, e.g., "*In <u>monsoon regions</u> <u>rainfall</u> does not* …". He further could not unambiguously classify as left or right bracketed another 35 examples (or 11%), e.g., "*Most advanced aircraft have <u>precision navigation systems</u>.*". After having discarded these two categories, he was left with 244 unambiguous three-word noun compounds to be used for testing: 163 left-bracketed and 81 right-bracketed. His algorithms are unsupervised and need no training data.

Table 3.3 shows some examples from this dataset. Note that the resulting *Lauer's dataset* contains repetitions: there are only 215 distinct examples out of all 244.

Lauer used his own left/right in-context judgments only. However, he also performed an experiment with seven additional human annotators, who were asked to annotate the three-word noun compounds alone, *out-of-context*. Lauer split the 215 examples into

Noun Compound			Bracketing
minority	business	development	left
satellite	data	systems	right
disaster	relief	assistance	left
county	extension	agents	right
world	food	production	right
granary	storage	baskets	right
customs	enforcement	vehicles	left
airport	security	improvements	left
mountain	summit	areas	left
law	enforcement	agencies	left
college	commencement	poem	right
health	education	institutions	left
country	music	theme	left
sea	transportation	hub	left
army	ant	behavior	left
missile	defense	systems	left
world	petroleum	production	right
arab	independence	movements	left
speech	recognition	system	left
production	passenger	vehicles	right
revenue	ton	miles	right
combustion	chemistry	technology	left
science	fiction	novels	left
missile	guidance	system	left
sea	bass	species	left
radiation	energy	conversion	left
science	fiction	novels	left
energy	distribution	properties	left
science	fiction	writer	left
science	fiction	themes	left
breeder	technology	development	left
landmark	majority	opinions	right
community	college	system	left
town	council	members	left
war	crimes	prosecutor	left
health	insurance	laws	left
science	fiction	satire	left
death	penalty	statutes	left
calvinist	peasant	family	right
exhibition	ballroom	dancers	right

Table 3.3: **Lauer's dataset:** sample noun compounds and bracketing labels.

two blocks, with three judges assigned to the first block and four judges to the second block. On average the judges' judgments matched Lauer's annotations 81.5% of the time, and any two judges agreed between themselves on between 73% and 84% of the instances.

Note that this 81.5% agreement cannot be considered as an "upper bound" for his the test set: the actual test set contains 244 examples (with the repetitions), while the inter-annotator agreement was measured over 215 examples only. In addition, Lauer's programs had access to a text corpus and to a thesaurus, while the human judges were given the noun compounds only, in isolation. Therefore, it should not be surprising if some computer system manages to beat the 81.5% "upper bound" on accuracy.

3.8.2 Biomedical Dataset

I assembled a new dataset from the biomedical literature: the *Biomedical dataset*. Unlike *Lauer's dataset*, this dataset is extracted using a more sophisticated language technology, from a much larger collection (in the biomedical domain), and has been annotated by two independent judges. See section C.1 in appendix C for a detailed description of the process of extraction of the most frequent three-word noun compounds in a 1.4 million abstracts subset of MEDLINE.

The top 500 extracted noun compounds are annotated independently by two judges: myself and a student with a biological background. The problematic cases are reconsidered by the two judges and, after agreement is reached, the set contains 361 left-bracketed, 69 right-bracketed, and 70 ambiguous noun compounds. The latter group is not used in the evaluation. The 430 unambiguous examples are shown in section C.3 in appendix C: the 70 ambiguous examples are omitted since they are not used in the evaluation.

The inter-annotator agreement for the original annotator choices is 82%. In addition, I calculate the value of the Kappa statistics, which measures the agreement between two (Cohen 1960) or more (Fleiss 1981) annotators on categorial judgments, taking into account the probability of agreement by chance:

$$K = \frac{\Pr(A) - \Pr(E)}{1 - \Pr(E)} \tag{3.19}$$

where $\Pr(A)$ is the observed agreement, and $\Pr(E)$ is the agreement by chance.

K is 1 in case of complete agreement and is 0 if the agreement matches what is expected by chance. The statistics can be less than 0 as well if the agreement is less than the expected by chance. The values of K can be interpreted as follows (Siegel & Jr. 1988):

- 0.00–0.20: slight agreement;
- 0.21–0.40: fair agreement
- 0.41–0.60: moderate agreement;
- 0.61–0.80: substantial agreement;
- 0.81–1.00: (almost) perfect agreement.

For the *Biomedical dataset* the value of the Kappa statistics was 0.442, which corresponds to moderate inter-annotator agreement. If we consider correct the examples marked as ambiguous by one of the annotators only, the kappa statistics becomes .606, a substantial agreement.

3.9 Evaluation

3.9.1 Experimental Setup

I experimented with *Lauer's dataset* (see section 3.8.1) in order to make my results comparable to those of previous researchers, e.g., Lauer (1995), Lapata & Keller (2004), etc. I also experimented with the *Biomedical dataset*, described in section 3.8.2 above.

I collected statistics for the n-grams, the surface features, and the paraphrases by issuing exact phrase queries, limiting the pages to English and requesting filtering of the similar results. I used *MSN Search* (now *Live Search*) statistics for the n-grams and the paraphrases (unless the pattern contains an asterisk, in which case I used *Google*), and *Google* for the surface features. *MSN Search* always returns exact numbers as page hits estimates, while *Google* and *Yahoo!* return rounded estimates. For example, at the time of writing, *Live Search* returns 60,219,609 page hits for *cell*, while *Google* and *Yahoo!* return 397,000,000 and 637,000,000, respectively. This rounding can be problematic, especially when comparing two ratios of page hits, e.g., $\frac{\#(brain,stem)}{\#(stem)}$ and $\frac{\#(stem,cell)}{\#(cell)}$.

For each noun compound, I generated all possible word inflections (e.g., *tumor* and *tumors*) and alternative word variants (e.g., *tumor* and *tumour*). For this purpose, for *Lauer's dataset*, I used *Carroll's morphological tools*[12] (Minnen *et al.* 2001) and the *Java WordNet Library*[13] (*JWNL*), which provides programmatic access to *WordNet* (Fellbaum 1998). For the *Biomedical dataset*, I also used the *UMLS Specialist Lexicon*.[14]

For bigrams, I inflected the second word only. Similarly, for a prepositional para-

[12] `http://www.cogs.susx.ac.uk/lab/nlp/carroll/morph.html`
[13] `http://sourceforge.net/projects/jwordnet`
[14] `http://www.nlm.nih.gov/pubs/factsheets/umlslex.html`

phrase, I generated the inflected forms for the two parts, before and after the preposition. See sections 3.4 and 3.5 for a detailed description of what kinds of inflections are used for each model and how the page hits are summed up.

3.9.2 Results for *Lauer's Dataset*

The results of the evaluation for *Lauer's dataset* are shown in Table 3.4. Note that the results for the accuracy are accompanied by confidence intervals, which will be explained in section 3.9.4 below. The left-bracketing baseline model yields 66.8% accuracy. The most accurate models are the *concatenations triple* (96.2%), *genitive* (88.89%) and *abbreviation* (87.5%), which only make predictions for about 10-30% of the examples. The *surface features* and the *paraphrases* achieve 85.51% and 82.08%, respectively, with almost 90% coverage. All these models outperform the word association models (*adjacency* and *dependency*: frequency, probability, PMI, χ^2), whose accuracy is below 80% (with 100% coverage). Among the word association models, the *dependency* models clearly outperform the *adjacency* models, which is consistent with what other researchers previously have reported (Lauer 1995; Lapata & Keller 2004). Using patterns containing *Google*'s operator performs worse than the dependency-based word association models: the coverage is in the eighties (as opposed to 100%), and there is a 1-5% absolute drop in accuracy.

The final bracketing decision is a majority vote combination of the predictions of the models shown in bold in Table 3.3: χ^2 *adjacency*, χ^2 *dependency*, *concatenation dependency*, *concatenation triples*, *genitive markers*, *abbreviations*, *paraphrases*, and *surface features*. This yields 90.52% accuracy and 95.08% coverage. Defaulting the unassigned cases to left bracketing yields 89.34% accuracy.

Model	**Correct**	**Wrong**	**N/A**	**Accuracy**	**Cover.**
# adjacency	183	61	0	75.00±5.79	100.00
Pr adjacency	180	64	0	73.77±5.86	100.00
PMI adjacency	182	62	0	74.59±5.81	100.00
χ^2 **adjacency**	184	60	0	**75.41±5.77**	100.00
# dependency	193	50	1	79.42±5.52	99.59
Pr dependency (= PMI dep.)	194	50	0	79.51±5.50	100.00
χ^2 **dependency**	195	49	0	**79.92±5.47**	100.00
# adjacency (*)	152	41	51	78.76±6.30	79.10
# adjacency (**)	162	43	39	79.02±6.08	84.02
# adjacency (***)	150	51	43	74.63±6.44	82.38
# adjacency (*, rev.)	163	48	33	77.25±6.11	86.47
# adjacency (**, rev.)	165	51	28	76.39±6.09	88.52
# adjacency (***, rev.)	156	57	31	73.24±6.32	87.30
Concatenation adjacency	175	48	21	78.48±5.85	91.39
Concatenation dependency	167	41	36	**80.29±5.93**	85.25
Concatenation triples	76	3	165	**96.20±6.78**	32.38
Inflection variability	69	36	139	65.71±9.49	43.03
Swap first two words	66	38	140	63.46±9.58	42.62
Reorder	112	40	92	73.68±7.52	62.30
Abbreviations	21	3	220	**87.50±18.50**	9.84
Genitive marker	32	4	208	**88.89±14.20**	14.75
Paraphrases	174	38	32	**82.08±5.72**	86.89
Surface features (sum)	183	31	30	**85.51±5.34**	87.70
Majority vote	210	22	12	90.52±4.46	95.08
Majority vote → 'left'	*218*	*26*	*0*	***89.34±4.50***	***100.00***
Baseline (choose 'left')	163	81	0	66.80±6.13	100.00

Table 3.4: **Bracketing results on *Lauer's dataset*.** For each model, the number of correctly classified, wrongly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% examples for which the model makes prediction).

Model	Accuracy (%)
baseline (*'left'*)	66.80±6.13
(Lauer 1995) adjacency	68.90±6.07
(Lauer 1995) dependency	77.50±5.65
My χ^2 dependency	*79.92±5.47*
(Lauer 1995) tuned	80.70±5.41
My majority vote → left	***89.34±4.50***
(Lapata & Keller 2004): baseline (122 examples)	63.93±8.83
(Lapata & Keller 2004): best *BNC* (122 examples)	68.03±8.72
(Lapata & Keller 2004): best *AltaVista* (122 examples)	78.68±8.08
⋆(Girju *et al.* 2005): best C5.0 (shuffled dataset)	83.10±5.20

Table 3.5: **Comparison to other unsupervised results on *Lauer's dataset*.** The results of Lapata & Keller (2004) are on half of *Lauer's dataset*: note the different baseline. The model of Girju *et al.* (2005) is supervised; it also mixes *Lauer's dataset* with additional data. See section 3.2 for more details.

Table 3.5 compares my results on *Lauer's dataset* to the results of Lauer (1995), Lapata & Keller (2004), and Girju *et al.* (2005). It is important to note that, while my results are *directly* comparable to those of Lauer (1995), the ones of Lapata & Keller (2004) are not since they use half of *Lauer's dataset* for development (122 examples) and the other half for testing[15]. Following Lauer (1995), I use the whole dataset for testing. In addition, the Web-based experiments of Lapata & Keller (2004) use the *AltaVista* search engine, which no longer exists in its earlier form: it has been acquired by *Yahoo!*.

Table 3.5 also shows the results of Girju *et al.* (2005), who achieve 83.1% accuracy, but use a *supervised* algorithm which targets bracketing *in context*. They further "shuffle" *Lauer's dataset*, mixing it with additional data, which makes it hard to compare directly to their results. More details can be found in section 3.2 above.

[15] However, the differences are negligible; their system achieves very similar results on the whole dataset (personal communication).

Model	**Correct**	**Wrong**	**N/A**	**Accuracy**	**Cover.**
# adjacency	374	56	0	86.98±3.51	100.00
Pr adjacency	353	77	0	82.09±3.90	100.00
PMI adjacency	372	58	0	86.51±3.55	100.00
χ^2 adjacency	379	51	0	**88.14±3.40**	100.00
# dependency	374	56	0	86.98±3.51	100.00
Pr dependency (= PMI dep.)	369	61	0	85.81±3.62	100.00
χ^2 dependency	380	50	0	**88.37±3.38**	100.00
# adjacency (*)	373	57	0	86.74±3.53	100.00
# adjacency (**)	358	72	0	83.26±3.82	100.00
# adjacency (***)	334	88	8	79.15±4.13	98.14
# adjacency (*, rev.)	370	59	1	86.25±3.58	99.77
# adjacency (**, rev.)	367	62	1	85.55±3.64	99.77
# adjacency (***, rev.)	351	79	0	81.63±3.93	100.00
Concatenation adjacency	370	47	13	88.73±3.40	96.98
Concatenation dependency	366	43	21	**89.49±3.35**	95.12
Concatenation triple	238	37	155	**86.55±4.54**	63.95
Inflection variability	198	49	183	80.16±5.42	57.44
Swap first two words	90	18	322	83.33±8.15	25.12
Reorder	320	78	32	80.40±4.18	92.56
Abbreviations	133	23	274	**85.25±6.41**	36.27
Genitive markers	48	7	375	**87.27±11.29**	12.79
Paraphrases	383	44	3	**89.70±3.25**	99.30
Surface features (sum)	382	48	0	**88.84±3.33**	100.00
Majority vote	403	17	10	95.95±2.34	97.67
Majority vote → 'right'	*410*	*20*	*0*	***95.35±2.42***	***100.00***
Baseline (choose 'left')	361	69	0	83.95±3.77	100.00

Table 3.6: **Bracketing results on the *Biomedical dataset* using the Web.** For each model, the number of correctly classified, wrongly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% examples for which the model makes prediction).

3.9.3 Results for the *Biomedical Dataset*

The results for the *Biomedical dataset* are shown in Table 3.6. This dataset has a high left-bracketing baseline of almost 84% accuracy. The *adjacency* and the *dependency* word association models yield very similar accuracy; in both cases χ^2 is a clear winner with over 88%, outperforming the other word association models by 1.5-6%. Further, the *frequency* models outperform the *probability* models due to the abundance of words with unreliable Web frequency estimates: single-letter (e.g., *T cell*, *vitamin D*), Roman digits (e.g., *ii*, *iii*), Greek letters (e.g., *alpha*, *beta*), etc. These are used by the *probability* model, but not by the *frequency* model, which uses bigrams only. While being problematic for most models, these words work well with *concatenation dependency* (89.49% accuracy, 95.12% coverage), e.g., *T cell* often can be found written as *Tcell*.

As before, the different models are combined in a majority vote, using the same voting models as for *Lauer's dataset* above, which yields 95.95% accuracy and 97.67% coverage. Defaulting the unassigned cases to right bracketing yields 95.35% accuracy overall.

Finally, Table 3.7 shows the performance of the surface features on the *Biomedical dataset*. We can see that most of the assumed right-predicting surface features actually correlate better with a left bracketing. Overall, the surface features are very good at predicting left bracketing, but are unreliable for the right-bracketed examples due to scope ambiguity e.g., in "*brain stem-cell*", *cell* could attach to *stem* only, in which case it would correctly predict a right bracketing; however, it could also target the compound *brain stem*, in case of left-bracketed compound.

Example	**Predicts**	**Accuracy (%)**	**Coverage (%)**
brain-stem cells	left	**88.22**	**92.79**
brain stem's cells	left	**91.43**	**16.28**
(brain stem) cells	left	96.55	6.74
brain stem (cells)	left	100.00	1.63
brain stem, cells	left	96.13	42.09
brain stem: cells	left	97.53	18.84
brain stem cells-death	left	80.69	60.23
brain stem cells/tissues	left	83.59	45.35
brain stem Cells	left	**90.32**	**36.04**
brain stem/cells	left	100.00	7.21
brain. stem cells	left	97.58	38.37
brain stem-cells	right	**25.35**	**50.47**
brain's stem cells	right	**55.88**	**7.90**
(brain) stem cells	right	46.67	3.49
brain (stem cells)	right	0.00	0.23
brain, stem cells	right	54.84	14.42
brain: stem cells	right	44.44	6.28
rat-brain stem cells	right	17.97	68.60
neural/brain stem cells	right	16.36	51.16
brain Stem cells	right	24.69	18.84
brain/stem cells	right	53.33	3.49
brain stem. cells	right	39.34	14.19

Table 3.7: **Surface features analysis (%s)**, run over the *Biomedical dataset*.

3.9.4 Confidence Intervals

Table 3.4 shows that the accuracy for the χ^2 *dependency* model is 79.92%; however it depends on the particular testing dataset used in the evaluation: *Lauer's dataset*. What if I repeated the experiments with another dataset of 244 examples extracted in the same way from the same data source, i.e. from *Grolier's encyclopedia*? Probably, accuracy values like 79.6% or 81.3% should not be surprising, while 50% would be extremely rare. The question then becomes: Which values are to be expected and which ones should be considered particularly unusual?

The most popular statistical answer to this question is to provide a whole interval of likely values, rather than a single estimate, and the most widely used frequentist approach is to construct a *confidence interval*, which is an interval of values calculated in a way so that if the experiment is repeated with multiple samples (here datasets) from the same population, the calculated confidence interval (which would be different for each sample) would contain the true population parameter (here the accuracy for χ^2 *dependency*, if all possible three-word noun compounds from *Grolier's encyclopedia* were to be tested) for a high proportion of these samples.

More formally, given an unobservable parameter θ, the random interval (U, V) is called a *confidence interval for* θ, if U and V are observable random variables whose probability distribution depends on θ, and $\Pr(U < \theta < V|\theta) = x$. The number x $(0 \leq x \leq 1)$ is called *confidence level* and is usually given in per cents: the value of 95% is typically used, although 90%, 99% and 99.9% are common as well, with a higher confidence level corresponding to a wider interval.

For example, given a sample mean $\hat{\mu}$, the confidence interval for the true mean μ of a normally-distributed random variable whose population's standard deviation σ_μ is known can be calculated as follows:

$$\hat{\mu} \pm z_{(1-\alpha/2)}\sigma_\mu \tag{3.20}$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ percentile of the standard normal distribution. For a 95% confidence level, we set $\alpha = 0.05$.

In my case, I need to calculate a confidence interval for the accuracy, which is a *confidence interval for a proportion*: the proportion of correct examples out of all classified examples. Since there is a fixed number of examples (244 for *Lauer's dataset*), with two possible outcomes for each one (correct and wrong), assuming statistically independent trials and the same probability of success for each trial, a *binomial distribution* can be assumed. Using the *central limit theorem*, this binomial distribution can approximated with a normal distribution, which yields the *Wald interval* (Brown *et al.* 2001):

$$\hat{p} \pm z_{(1-\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3.21}$$

where n is the sample size, and $\hat{p}$ is an estimation for the proportion.

Here is an example: According to Table 3.4, the χ^2 *dependency* model makes 195 correct and 49 wrong classification decisions. Therefore, $n = 195 + 49 = 244$ and $\hat{p} = 195/244 = 0.7992$. A 95% confidence level corresponds to $\alpha = 0.05$, and thus $z_{(1-\alpha/2)}$ is $z_{0.975}$, which is found in a statistical table for the standard normal distribution to be 1.96. Substituting these values in eq. 3.21 yields 0.7992 ± 0.05. Therefore, the 95% confidence Wald interval for χ^2 *dependency* model's accuracy is [0.7492, 0.8492] or [74.92%, 84.92%].

While the Wald interval is simple to calculate, there is a more accurate alternative, the *Wilson interval* (Wilson 1927):

$$\frac{\hat{p} + \frac{1}{2n} z^2_{(1-\alpha/2)} \pm z_{(1-\alpha/2)} \sqrt{\frac{1}{n}[\hat{p}(1-\hat{p}) + \frac{1}{4n} z^2_{(1-\alpha/2)}]}}{1 + \frac{1}{n} z^2_{(1-\alpha/2)}} \tag{3.22}$$

For the χ^2 *dependency* model, the above formula yields 79.92% ± 5.47%, which is the interval shown in Table 3.4. Because of its higher accuracy (Agresti & Coull 1998), I use the Wilson interval for all calculations of confidence intervals for proportions in this chapter, as well as elsewhere in the thesis, unless otherwise stated.

3.9.5 Statistical Significance

Table 3.5 shows that the accuracy for my χ^2 *dependency* model is 79.92%, while for Lauer's (1995) dependency model it is 77.5%. Apparently, the former model is better than the latter one, but maybe this is due to chance alone? What about the almost 9% difference in accuracy between my "*Majority vote* → *left*" (89.34% accuracy) and Lauer's (1995) tuned model (80.7% accuracy): could it have occurred by chance too?

These are questions about *statistical significance*: an outcome is considered to be *statistically significant*, if it is unlikely to have occurred due to chance alone. The statistical significance of a result is characterized by its p-value, which is the probability of obtaining an outcome at least as extreme as the observed one. A test for statistical significance calculates a p-value (under some assumptions) and then compares it to a pre-specified *significance level* α (e.g., 0.05): the result is considered statistically significant, if the p-value is less than α.

	Correct	**Wrong**
Model 1	A	B
Model 2	C	D

Table 3.8: **The contingency table:** used in a χ^2 test for testing if the difference in performance between two models is statistically significant.

There are various statistical tests depending on the kinds of observed events and on the assumptions about them that could be made. In particular, the most widely used test for checking whether the difference in performance between two models is statistically significant is the *Pearson's* χ^2 *test*.[16] The test makes use of a 2×2 *contingency table* like the one shown in Table 3.8: the rows are associated the models and the columns with the number of correctly/wrongly classified examples. It tries to determine whether the probability of an example being classified correctly or wrongly depends on the model used, i.e. it checks whether the null hypothesis, that the columns of the table are independent of the rows, can be rejected with a significance level α.

Let us consider for example, how the *Pearson's* χ^2 *test* can be applied to comparing the χ^2 *dependency* model (Model 1) and the Lauer's (1995) dependency model (Model 2), whose accuracies are 79.92% and 77.5%, respectively. Since there are 244 testing examples in total, we have $A = 189$, $B = 55$, $C = 195$, $D = 49$, as shown in Table 3.9. Let N be the total sum of all table entries, i.e. $N = A+B+C+D = 488$. Then the marginal probability of an example in this table being classified correctly regardless of the model is:

$$\Pr(correct) = \frac{A+C}{N} = \frac{189+195}{488} = 0.7869 \tag{3.23}$$

[16] Some other tests are suitable for the purpose as well, e.g., the *Fisher's exact test*.

	Correct	Wrong
Model 1	$A = 189$ (192)	$B = 55$ (52)
Model 2	$C = 195$ (192)	$D = 49$ (52)

Table 3.9: **Sample contingency table:** the values in parentheses are the expected values under the null hypothesis that the rows and the columns are independent.

Similarly, the marginal probability of an example in this table being classified by Model 1 is:

$$\Pr(Model1) = \frac{A+B}{N} = \frac{189+55}{488} = 0.5 \tag{3.24}$$

Under the null hypothesis that the rows and the columns of the table are independent, the probability of an example being classified correctly by Model 1 would be the product of the above marginal probabilities:

$$\Pr(correct, Model1) = \Pr(correct) \times \Pr(Model1) \tag{3.25}$$

Therefore, the expected number of examples being classified correctly by Model 1 would be:

$$E(correct, Model1) = N \times \Pr(correct, Model1) = 192 \tag{3.26}$$

The expected values for the remaining three table entries can be calculated in a similar manner:

$$E(correct, Model2) = N \times \Pr(correct, Model2) = 192 \tag{3.27}$$

$$E(wrong, Model1) = N \times \Pr(correct, Model2) = 52 \tag{3.28}$$

$$E(wrong, Model2) = N \times \Pr(correct, Model2) = 52 \tag{3.29}$$

These expected values for the table entries are shown in parentheses in Table 3.9. The χ^2 score is calculated using the following formula:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{3.30}$$

where O_i are the observed values (not in parentheses in Table 3.9) and E_i are the expected values (shown in parentheses in Table 3.9). The summation is over all table entries.

Substituting the values in Table 3.9 in the above formula (3.30), we obtain:

$$\chi^2 = \frac{(189 - 192)^2}{192} + \frac{(55 - 52)^2}{52} + \frac{(195 - 192)^2}{192} + \frac{(49 - 52)^2}{52} \tag{3.31}$$

I then consult a table of the χ^2 distribution with 1 degree of freedom[17], and I find that the probability of observing this or a more extreme difference, under the null hypothesis is approximately 0.5072, i.e. the p-value equals 0.5072. Since 0.5072 is quite a big probability, much bigger than $\alpha = 0.05$, the null hypothesis cannot be rejected.

As I already mentioned in section 3.4.4, there is more direct way to calculate the χ^2 score from a 2×2 table (repeated here for completeness):

$$\chi^2 = \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{3.32}$$

What about Lauer's (1995) tuned model and my "*Majority vote* $\rightarrow$ *left*" model? The accuracies are 80.7% and 89.34%, respectively, and since there are a total of 244 examples then $A = 197$, $B = 47$, $C = 218$, $D = 26$. The χ^2 score is 7.104, and the corresponding p-value equals 0.0077, which means that the null hypothesis can be rejected with the standard significance level $\alpha = 0.05$, and even with the much lower $\alpha = 0.01$. The difference between these models is considered very statistically significant.

[17] One degree of freedom, since there are two outcomes, *correct* and *wrong*, whose sum is fixed.

Is the difference between my "*Majority vote* $\rightarrow$ *left*" and the best previously published result, the supervised C5.0 model of Girju *et al.* (2005) significantly different? The corresponding accuracies are 89.34% and 83.1%, respectively, and thus $A = 218$, $B = 26$, $C = 203$, and $D = 41$. The χ^2 score equals 3.893, which corresponds to a p-value of 0.0485, and therefore, the difference is statistically significant with a significance level $\alpha = 0.05$. Thus, I can conclude that my result is statistically significantly better (with a significance level $\alpha = 0.05$), on *Lauer's dataset* or a variation of it, than any previously proposed algorithm: both supervised and unsupervised.

Note that in the last calculation the evaluation datasets are different (but overlapping): the dataset of Girju *et al.* (2005) still contains 244 examples, but some of them are not from the original dataset (see section 3.2). This is not a problem: the χ^2 test is robust against such differences, provided that the distribution does not change. The test is also robust against different numbers of examples, which allows me to compare my results or those of Lauer (1995), which use all of *Lauer's dataset*, with the ones of Lapata & Keller (2004), who only use half of it.

Finally, there is a connection between confidence intervals and statistical significance: a confidence interval around a particular value includes all other values that are not statistically significantly different from it. For example, as I have calculated in section 3.9.4 above, the 95% confidence interval around the 79.92% value for the accuracy of the χ^2 *dependency model* is [74.92%, 84.92%]. Therefore, the 80.7% accuracy for Lauer's (1995) tuned model is not statistically significantly different from 79.92% since it is in the interval, but the value of 89.34% for my combined model is since it is outside of the interval.

Model	Correct	Wrong	N/A	Accuracy	Cover.
# adjacency	196	36	0	84.48±5.22	100.00
Pr adjacency	173	59	0	74.57±5.97	100.00
χ^2 adjacency	200	32	0	86.21±5.03	100.00
# dependency	195	37	0	84.05±5.26	100.00
Pr dependency	193	39	0	83.19±5.34	100.00
χ^2 dependency	196	36	0	84.48±5.22	100.00
PrepPar	181	13	38	93.30±4.42	83.62
PP+χ^2adj+χ^2dep	207	13	12	94.09±3.94	94.83
PP+χ^2adj+χ^2dep→right	214	18	0	**92.24±4.17**	100.00
Baseline (choose left)	193	39	0	83.19±5.34	100.00

Table 3.10: **Bracketing results on the *Biomedical dataset* using 1.4M MEDLINE abstracts.** For each model, the number of correctly classified, wrongly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (in %).

3.10 Using MEDLINE instead of the Web

For comparison purposes, I experimented with the *Biomedical dataset* using a domain-specific text corpus with suitable linguistic annotations instead of the Web. I use the Layered Query Language and architecture, described in section C.1 in appendix C, in order to acquire n-gram and paraphrase frequency statistics.

My corpus consists of about 1.4 million MEDLINE abstracts, each one being about 300 words long on the average, which means about 420 million indexed words in total. For comparison, *Google* indexes about eight billion pages; if we assume that each one contains about 500 words on the average, this yields about four trillion indexed words, which is about a million times bigger than my corpus. Still, the subset of MEDLINE I use is about four times bigger than the 100 million word *BNC* used by Lapata & Keller (2004). It is also more than fifty times bigger than the eight million word *Grolier's encyclopedia* used by Lauer (1995).

The queries I used to collect n-gram and paraphrases counts are described in section C.2 in appendix C. The results are shown in Table 3.10. In addition to probabilities (Pr), I also use counts (#) and χ^2 (with the dependency and the adjacency models). The prepositional paraphrases are much more accurate: 93.3% (with 83.62% coverage). By combining the paraphrases with the χ^2 models in a majority vote, and by assigning the undecided cases to right-bracketing, I achieve 92.24% accuracy, which is slightly worse than 95.35% I achieved using the Web. This difference is not statistically significant[18], which suggests that in some cases a big domain-specific corpus with suitable linguistic annotations could be a possible alternative of the Web. This is not true, however, for general domain compounds: for example, my subset of MEDLINE can provide prepositional paraphrases for only 23 of the 244 examples in *Lauer's dataset* (i.e. for less than 10%), and for 12 of them the predictions are wrong (i.e., the accuracy is below 50%).

3.11 Conclusion and Future Work

I have described a novel, highly accurate lightly supervised approach to noun compound bracketing which makes use of novel surface features and paraphrases extracted from the Web and achieves a statistically significant improvement over the previous state-of-the-art results for *Lauer's dataset*. The proposed approach is more robust than the one proposed by Lauer (1995) and more accurate than that of Lapata & Keller (2004). It does not require labeled training data, lexicons, ontologies, thesauri, or parsers, which makes it promising for a wide range of other NLP tasks.

[18]Note however that here I experiment with 232 of the 430 examples.

A simplification of the method has been used by Vadas & Curran (2007) in order to augment the NPs in the *Penn Treebank* with internal structure. Some of the proposed features have been found useful for query segmentation. by Bergsma & Wang (2007).

An important direction for future work is in reducing the number of queries to the search engine. One solution would be to perform a careful analysis and to exclude the most expensive patterns and those with a minor positive contribution to the overall accuracy. Using *Google*'s *Web 1T 5-gram* dataset of 5-gram that appear on the Web at least 40 times and their corresponding frequencies is another promising direction to go. However, it might be of limited use since many of my patterns are longer than 5-grams and/or appear less than 40 times on the Web.

Chapter 4

Noun Compounds' Semantics and Relational Similarity

In this chapter, I present a novel, simple, unsupervised method for characterizing the semantic relations that hold between nouns in noun-noun compounds. The main idea is to look for *predicates* that make explicit the hidden relations between the nouns. This is accomplished by writing Web search engine queries that restate the noun compound as a relative clause containing a wildcard character to be filled in with a verb. A comparison to results from the literature and to human-generated verb paraphrases suggests this is a promising approach. Using these verbs as features in classifiers, I demonstrate state-of-the-art results on various relational similarity problems: mapping noun-modifier pairs to abstract relations like `TIME`, `LOCATION` and `CONTAINER`, classifying relation between nominals, and solving SAT verbal analogy problems. A preliminary version of some of these ideas appeared in (Nakov & Hearst 2006) and (Nakov & Hearst 2007).

4.1 Introduction

While currently there is no consensus as to what relations can hold between nouns in a noun compound, most proposals make use of small sets of abstract relations, typically less than fifty. Some researchers, however, have proposed that an unlimited number of relations is needed (Downing 1977). I hold a similar position.

Many algorithms that perform semantic interpretation place heavy reliance on the appearance of verbs, since they are the predicates which act as the backbone of the assertion being made. Noun compounds are terse elisions of the predicate; their structure assumes that the reader knows enough about the constituent nouns and about the world at large to be able to infer what the relationship between the words is. Here I propose to try to uncover the relationship between the noun pairs by, in essence, rewriting or paraphrasing the noun compound in such a way as to be able to determine the predicate(s) holding between the nouns. Therefore, I represent noun compounds semantics in terms of verbs, rather than a fixed number of abstract predicates (Levi 1978) (e.g., `HAVE`, `MAKE`, `USE`), relations (Girju *et al.* 2005) (e.g., `LOCATION`, `INSTRUMENT`, `AGENT`), or prepositions (Lauer 1995) (e.g., `OF`, `FOR`, `IN`), as is traditional in the literature. The idea is similar to the approach of Finin (1980), who characterizes the relation in a noun-noun compound using an inventory of all possible verbs that can link the noun constituents, e.g., *salt water* is interpreted using relations like *dissolved in*.

In the proposed approach, I pose paraphrases for a given noun compound by rewriting it as a phrase that contains a wildcard where a verb would go. For example, I rewrite *neck vein* as `"vein that * neck"`, I send this as a query to a Web search engine, and then I parse the resulting snippets in order to find the verbs that appear in the place of the wildcard. For example, the most frequent verbs (+ prepositions) I find for *neck vein* are: *emerge from*, *pass through*, *be found in*, *be terminated at*, *be in*, *run from*, *descend in*.

4.2 Related Work

The syntax and semantics of noun compounds are active areas of research, which are part of the broader research on multi-word expressions (MWEs): there have been workshops on MWEs as part of the annual meetings of the Association for Computational Linguistics (ACL) in 2003, 2004, 2006 and 2007, and during the meeting of the European chapter of ACL (EACL) in 2006. In 2007, the SemEval workshop on Semantic Evaluations (formerly *Senseval*), co-located with the annual meeting of the ACL, had a specialized task on *Classification of Semantic Relations between Nominals* (Girju *et al.* 2007). Finally, there was a special issue on MWEs of the *Journal of Computer Speech and Language* in 2005 (Villavicencio *et al.* 2005), and there are two upcoming special issues of the *International Journal of Language Resources and Evaluation* in 2008: one on *Multi-word Expressions*, and another one on '*Computational Semantic Analysis of Language: SemEval-2007 and Beyond*'.

Lauer (1995) defines the semantic relation identification problem as one of predicting which among the following eight prepositions is most likely to be associated with the noun compound when rewritten: *of*, *for*, *in*, *at*, *on*, *from*, *with* and *about* (see Table 2.3). This approach can be problematic since the semantics of the prepositions is vague, e.g., *in*, *on*, and *at*, all can refer to both `LOCATION` and `TIME`. Lauer builds a corpus-based model for predicting the correct preposition, and achieves 40% accuracy. Lapata & Keller (2005) improve on these results the Web to estimate trigram frequencies for ($noun_1$, $prep$, $noun_2$), achieving 55.71% accuracy.

Rosario & Hearst (2001) use their own dataset of 18 abstract semantic classes and show that a discriminative classifier can work quite well at assigning relations from a pre-defined set if training data is supplied in a domain-specific setting: 60% accuracy.

Rosario *et al.* (2002) report 90% accuracy using a simple “descent of hierarchy” approach, which characterizes the relation between two nouns in a bioscience noun-noun compound based on the semantic category each of the constituent nouns belongs to. See section 2.5.3 for a detailed description.

Girju *et al.* (2004) present an SVM-based approach for the automatic classification of semantic relations in nominalized noun phrases, where either the head or the modifier has been derived from a verb. Their classification schema consists of 35 abstract semantic relations and has been also used by Moldovan *et al.* (2004) for the semantic classification of noun phrases in general, including complex nominals, genitives and adjectival noun phrases.

Girju *et al.* (2005) apply both classic (SVM and decision trees) and novel supervised models (semantic scattering and iterative semantic specialization), using *WordNet*, word sense disambiguation, and a set of linguistic features. They test their system against both Lauer’s eight prepositional paraphrases and their own set of 21 semantic relations, achieving up to 54% accuracy on the latter.

Girju (2007b) uses cross-linguistic evidence from a set of five Romance languages, from which NP features are extracted and used in an SVM classifier, which yields 77.9%

accuracy when *Europarl* is used, and 74.31% with *CLUVI*. Additional details on the features and on the process of their extraction are provided in (Girju 2007a).

Lapata (2002) focuses on the disambiguation of nominalizations – a particular class of noun compounds whose head is derived from a verb and whose modifier was an argument of that verb. Using partial parsing, sophisticated smoothing and contextual information, she achieves 86.1% accuracy (baseline 61.5% accuracy) on the binary decision task of whether the modifier used to be the subject or the object of the nominalized verb, i.e. the head.

Kim & Baldwin (2006) try to characterize the semantic relation in a noun-noun compound using the verbs connecting the two nouns by comparing them to a predefined set of seed verbs, using a memory-based classifier (TiMBL). There are two problems with their approach: it is highly resource intensive (uses *WordNet*, *CoreLex* and *Moby's thesaurus*), and it is quite sensitive to the seed set of verbs: on a collection of 453 examples and 19 relations, they achieve 52.6% accuracy with 84 seed verbs, but only 46.66% when only 57 seed verbs were used.

Kim & Baldwin (2007b) describe a bootstrapping method for automatically tagging noun compounds with their corresponding semantic relations. Starting with 200 seed annotated noun compounds they replaced one constituent of each noun compound with similar words that are derived from synonyms, hypernyms and sister words, achieving accuracy ranging between 64.72% and 70.78%.

Kim & Baldwin (2007a) apply word sense disambiguation techniques to help supervised and unsupervised noun compound interpretation.

Séaghdha & Copestake (2007) use grammatical relations as features with an SVM classifier to characterize a noun-noun compound.

Devereux & Costello (2006) propose a vector-space model as representation model for the relations used to interpret noun-noun compounds. They use a fixed set of 19 head-modifier (H-M) relations like *H causes M*, *M causes H*, *H for M*, *H by M*, *M is H*. The meaning of a noun compound is hypothesized to be characterized by distribution over these 19 dimensions, as opposed to be expressed by a single relation. The model builds on the CARIN theory of conceptual combination (Gagné 2002) and was evaluated by measuring people's reaction time.

Turney *et al.* (2003) describe 13 independent classifiers, whose predictions are weighted and combined in order to solve SAT verbal analogy problems. They assembled a dataset of 374 questions, which were subsequently used by other researchers, e.g., Veale (2004), who achieved 43% accuracy on the same problem, using *WordNet* to calculate the similarity between individual words involved in the target SAT analogy problem instance.

Turney & Littman (2005) introduce a vector space model (VSM), and characterize the relation between two words, X and Y, as a fixed-length vector whose coordinates correspond to Web frequencies for 128 phrases like "X for Y", "Y for X", etc., derived from a fixed set of 64 joining terms (e.g. "for", "such as", "not the", "is *", etc.). These vectors are then used in a nearest-neighbor classifier to solve SAT analogy problems, achieving 47% accuracy (random-guessing baseline is 20%). In that paper they have applied that approach to classifying noun-modifier pairs into a fixed set of relations. Using a dataset of

600 examples created by Barker & Szpakowicz (1998), with 30 target relations they achieved an F-value of 26.5% (random guessing: 3.3%; majority-class baseline: 8.17%), and 43.2% when these 30 relations have been grouped into 5 course-grained relations (random guessing: 20%; majority-class baseline: 43.33%).

Turney (2006a) presents an unsupervised algorithm for mining the Web for patterns expressing implicit semantic relations. For example, `CAUSE` (e.g., *cold virus*) is best characterized by "*Y * causes X*", and the best pattern for `TEMPORAL` (e.g., *morning frost*) is "*Y in * early X*". This approach yields 50.2% F-value with 5 classes.

Turney (2005) introduces the latent relational analysis (LRA), which extends the VSM model of Turney & Littman (2005) by making use of automatically generated synonyms, by automatically discovering useful patterns, and by using a singular value decomposition in order to smooth the frequencies. The actual algorithm is quite complex and consists of 12 steps, described in detail in (Turney 2006b). When applied to the 374 SAT questions, it achieves the state-of-the-art accuracy of 56%. On the Barker & Szpakowicz (1998) dataset, the achieves an accuracy of 39.8% with 30 classes, and 58% with 5 classes.

Nulty (2007) uses a vector-space model representation where the vector coordinates are a fixed set of 28 joining terms like *of*, *for*, *from*, *without*, *across*, etc. The values of the coordinates are filled using Web n-gram frequencies. Using an SVM, he achieves 50.1% accuracy ona 20-fold cross-validation for the 5-class Barker&Szpakowicz dataset.

Most other approaches to noun compound interpretation use hand-coded rules for at least one component of the algorithm (Finin 1980), or rules combined with lexical resources (Vanderwende 1994) (52% accuracy, 13 relations). Barker & Szpakowicz (1998) make use of the identity of the two nouns and a number of syntactic clues in a nearest-neighbor classifier with 60-70% accuracy.

4.3 Using Verbs to Characterize Noun-Noun Relations

Traditionally the semantics of a noun compound have been represented as an abstract relation drawn from a small closed set. This is problematic for several reasons. First, it is unclear which is the best set, and mapping between different sets has proven challenging (Girju *et al.* 2005). Second, being both abstract and limited, such sets capture only part of the semantics; often multiple meanings are possible, and sometimes none of the pre-defined meanings are suitable for a given example. Finally, it is unclear how useful the proposed sets are, since researchers have often fallen short of demonstrating practical uses.

I believe verbs have more expressive power and are better tailored for the task of semantic representation: there is an infinite number of them and they can capture fine-grained aspects of the meaning. For example, while *wrinkle treatment* and *migraine treatment* express the same abstract relation `TREATMENT-FOR-DISEASE`, some fine-grained differences can be shown by specific verbs, e.g., *smooth* can paraphrase the former, but not the latter.

In many theories, verbs play an important role in the process of noun compound derivation (Levi 1978), and speakers frequently use them in order to make the hidden relation overt. This allows for simple extraction, but also for straightforward uses of the verbs and paraphrases in NLP tasks like machine translation, information retrieval, etc.

I further believe that a single verb often is not enough and that the meaning is approximated better by a collection of verbs. For example, while *malaria mosquito* can very well be characterized as `CAUSE` (or *cause*), further aspects of the meaning, can be captured by adding some additional verbs, e.g., *carry*, *spread*, *be responsible for*, *be infected with*, *transmit*, *pass on*, etc.

4.4 Method

In a typical noun-noun compound "$noun_1$ $noun_2$", $noun_2$ is the head and $noun_1$ is a modifier, attributing a property to it. The main idea of the proposed method is to preserve the head-modifier relation by substituting the pre-modifier $noun_1$ with a suitable post-modifying relative clause, e.g., "*tear gas*" can be transformed into "*gas that causes tears*", "*gas that brings tears*", "*gas which produces tears*", etc. Using all possible inflections of $noun_1$ and $noun_2$ from *WordNet*, I issue exact phrase *Google* queries of the following type:

```
"noun2 THAT * noun1"
```

where `THAT` is one of the following complementizers: *that*, *which* or *who*. The *Google* `*` operator stands for a one-word wildcard substitution; I issue queries with up to eight stars.

I collect the text snippets (summaries) from the search results pages (up to 1,000 per query) and I only keep the ones for which the sequence of words following `noun1` is non-empty and contains at least one non-noun, thus ensuring the snippet includes the entire noun phrase. To help POS tagging and shallow parsing the snippet, I further substitute the part before `noun2` by the fixed phrase "*We look at the*". I then perform POS tagging (Toutanova & Manning 2000) and shallow parsing[1], and I extract all verb forms, and the following preposition, if any, between `THAT` and `noun1`. I allow for adjectives and participles

[1]OpenNLP tools: `http://opennlp.sourceforge.net`

to fall between the verb and the preposition, but not nouns; I ignore the modal verbs and the auxiliaries, but I retain the passive *be*, and I make sure there is exactly one verb phrase (thus disallowing complex paraphrases like "*gas that makes the eyes fill with tears*"). Finally, I lemmatize the main verb using *WordNet*.

The proposed method is similar to previous paraphrase acquisition approaches which look for similar endpoints and collect the intervening material. Lin & Pantel (2001) extract paraphrases from dependency tree paths whose end points contain similar sets of words by generalizing over these ends, e.g., for "*X solves Y*", they extract paraphrases like "*X resolves Y*", "*Y is resolved by X*", "*X finds a solution to Y*", "*X tries to solve Y*", etc. The idea is extended by Shinyama *et al.* (2002), who use named entities of matching semantic classes as anchors, e.g., `LOCATION`, `ORGANIZATION`, etc. Unlike these approaches, whose goal is to create summarizing paraphrases, I look for verbs that can characterize noun compound semantics.

4.5 Semantic Interpretation

4.5.1 Verb-based Vector-Space Model

As an illustration of the method, consider the paraphrasing verbs (corresponding frequencies are shown in parentheses) it extracts for two noun-noun compounds with the same modifier and closely related synonymous heads: *cancer physician* and *cancer doctor*. Note the high proportion of shared verbs (underlined):

- **"cancer doctor"**: <u>*specialize in*</u>(12), <u>*treat*</u>(12), <u>*deal with*</u>(6), <u>*believe*</u>(5), <u>*cure*</u>(4), *attack*(4), <u>*get*</u>(4), *understand*(3), *find*(2), *miss*(2), *remove*(2), <u>*study*</u>(2), *know about*(2), *suspect*(2), *use*(2), *fight*(2), *deal*(2), *have*(1), *suggest*(1), *track*(1), *diagnose*(1), *recover from*(1), *specialize*(1), *rule out*(1), *meet*(1), *be afflicted with*(1), *study*(1), *look for*(1), *die from*(1), *cut*(1), *mention*(1), *cure*(1), *die of*(1), *say*(1), *develop*(1), *contract*(1)

- **"cancer physician"**: *specialize in*(11), *treat*(7), *have*(5), *diagnose*(4), *deal with*(4), *screen for*(4), *take out*(2), *cure*(2), *die from*(2), *experience*(2), *believe*(2), *include*(2), *study*(2), *misdiagnose*(1), *be treated for*(1), *work on*(1), *die of*(1), *survive*(1), *get*(1), *be mobilized against*(1), *develop*(1)

Now consider the following four different kinds of treatments:

- **"cancer treatment"**: *prevent*(8), *treat*(6), *cause*(6), *irradiate*(4), *change*(3), *help eliminate*(3), *be*(3), *die of*(3), *eliminate*(3), *fight*(3), *have*(2), *ask for*(2), *be specific for*(2), *decrease*(2), *put*(2), *help fight*(2), *die from*(2), *keep*(2), *be for*(2), *contain*(2), *destroy*(2), *heal*(2), *attack*(2), *work against*(2), *be effective against*(2), *be allowed for*(1), *stop*(1), *work on*(1), *reverse*(1), *characterise*(1), *turn*(1), *control*(1), *see*(1), *identify*(1), *be successful against*(1), *stifle*(1), *advance*(1), *pinpoint*(1), *fight against*(1), *burrow into*(1), *eradicate*(1), *be advocated for*(1), *counteract*(1), *render*(1), *kill*(1), *go with*(1)

- **"migraine treatment"**: *prevent*(5), *be given for*(3), *be*(3), *help prevent*(2), *help reduce*(2), *benefit*(2), *relieve*(1)

- **"wrinkle treatment"**: *reduce*(5), *improve*(4), *make*(4), *smooth*(3), *remove*(3), *be on*(3), *tackle*(3), *work perfect on*(3), *help smooth*(2), *be super on*(2), *help reduce*(2), *fight*(2), *target*(2), *contrast*(2), *smooth out*(2), *combat*(1), *correct*(1), *soften*(1), *reverse*(1), *resist*(1), *address*(1), *eliminate*(1), *be*(1)

- **"herb treatment"**: *contain*(19), *use*(8), *be concentrated with*(6), *consist of*(4), *be composed of*(3), *include*(3), *range from*(2), *incorporate*(1), *feature*(1), *combine*(1), *utilize*(1), *employ*(1)

Table 4.2 shows a subset of these verbs. As expected, *herb treatment*, which is quite different from the other compounds, shares no verbs with them: it *uses* and *contains* herb, but does not *treat* it. Further, while migraine and wrinkles cannot be *cured*, they can be *reduced*. Migraines can also be *prevented*, and wrinkles can be *smoothed*. Of course, these results are merely suggestive and should not be taken as ground truth, especially the absence of a verb. Still they seem to capture interesting fine-grained semantic distinctions, which normally require deep knowledge of the semantics of the two nouns and/or about the world.

The above examples suggest that paraphrasing verbs, and the corresponding frequencies, may be a good semantic representation from a computational linguistics point of view, e.g., they can be used in a vector space model in order to measure semantic similarity between noun-noun compounds.

I believe the paraphrasing verbs can be useful from a traditional linguistic (e.g., lexical semantics) point of view as well; I explore this idea below.

4.5.2 Componential Analysis

In lexical semantics, componential analysis is often used to represent the meaning of a word in terms of semantic primitives (features), thus reducing the word's meaning to series of binary components (Katz & Fodor 1963; Jackendoff 1983; Saeed 2003). For

	man	**woman**	**boy**	**bull**
ANIMATE	+	+	+	+
HUMAN	+	+	+	−
MALE	+	−	+	+
ADULT	+	+	−	+

Table 4.1: **Example componential analysis:** *man*, *woman*, *boy* and *bull*.

example, *bachelor* is (i) human, (ii) male, and (iii) unmarried, which can be expressed as [+HUMAN] [+MALE] [−MARRIED]. Similarly, *boy* can be analyzed as [+ANIMATE] [+HUMAN] [+MALE] [−ADULT]. See Table 4.1 for more examples.

Componential analysis has been very successful in phonology, where the sound system is limited and the contrast between different sounds is very important. For example, /p/ is distinguished from /b/ by the role of the vocal chords, and this distinction can be represented as a feature, e.g., /p/ is [−VOICED], while /b/ is [+VOICED].

In contemporary lexical semantics, componential analysis is considered useful for making explicit important semantic relations like hyponymy, incompatibility, etc., but is criticized for the following reasons: (1) it is unclear how the analysis decides on the particular features/components to include; and (2) it cannot really capture the full meaning of a given word.

4.5.3 Dynamic Componential Analysis

Given the similarity between Tables 4.1 and 4.2, I propose to analyze the semantics of the relations that hold between the nouns in a noun-noun compound using a kind of componential analysis, which I call *dynamic componential analysis*. The components of the

	cancer treatment	**migraine treatment**	**wrinkle treatment**	**herb treatment**
treat	+	+	+	−
prevent	+	+	−	−
cure	+	−	−	−
reduce	−	+	+	−
smooth	−	−	+	−
cause	+	−	−	−
contain	−	−	−	+
use	−	−	−	+

Table 4.2: **Some verbs characterizing different kinds of treatments.**

proposed model are paraphrasing verbs acquired dynamically from the Web in a principled manner, which addresses the major objection against the classic componential analysis of being inherently subjective.

4.6 Comparison to Abstract Relations in the Literature

In order to test the paraphrasing approach, I use noun-noun compound examples from the literature: I extract corresponding verbal paraphrases for them, and I manually determine whether these verbs accurately reflect the expected abstract semantic relations.

4.6.1 Comparison to Girju *et al.* (2005)

First, I study how my paraphrasing verbs relate to the abstract semantic relations proposed by Girju *et al.* (2005) for the semantic classification of noun compounds. For the purpose, I try to paraphrase the 21 example noun-noun compounds provided in that article as illustrations of the 21 abstract relations. I was only able to extract paraphrases for 14 of them, and I could not find meaningful verbs for the rest: *quality sound* (`ATTRIBUTE-HOLDER`), *crew investigation* (AGENT), *image team* (DEPICTION-DEPICTED), *girl mouth* (PART-WHOLE), *style performance* (MANNER), *worker fatalities* (RECIPIENT), and *session day* (MEASURE). For most of these seven cases, there appears either not to be a meaningful predicate for the particular nouns paired or a nominalization plays the role of the predicate.

Table 4.3 shows the target semantic relation, an example noun-noun compound from that relation, and the top paraphrasing verbs, optionally followed by prepositions, that I generate for that example. The verbs expressing the target relation are in bold, those referring to a different but valid relation are in italic, and the erroneous extractions are struck out. Each verb is followed by a corresponding frequency of extraction.

Overall, the extracted verbs seem to provide a good characterization of the noun compounds. While in two cases the most frequent verb is the copula (*to be*), the following most frequent verbs are quite adequate. In the case of *"malaria mosquito"*, one can argue that the CAUSE relation, assigned by Girju *et al.* (2005), is not exactly correct, in that the disease is only indirectly caused by the mosquitos (it is rather carried by them), and the proposed most frequent verbs *carry* and *spread* actually support a different abstract relation: AGENT. Still, *cause* appears as the third most frequent verb, indicating that it is common to consider indirect causation as a causal relation. In the case of *combustion gas*, the most frequent verb *support* is a good paraphrase of the noun compound, but is not directly applicable to the RESULT relation assigned by Girju *et al.* (2005); however, the remaining verbs for that relation do support RESULT.

For the remaining noun-noun compounds, the most frequent verbs accurately capture the relation assigned by Girju *et al.* (2005); in some cases, the less frequent verbs indicate other logical entailments for the noun combination.

4.6.2 Comparison to Barker & Szpakowicz (1998)

Table 4.4 compares my paraphrasing verbs and the first 8 (out of 20) abstract relations from Barker & Szpakowicz (1998): the paper gives several examples per relation, and I show the results for each of them, omitting *charitable donation* (`BENEFICIARY`) and *overdue fine* (`CAUSE`) since the modifier in these cases is an adjective[2], and *composer arranger* (`EQUATIVE`), for which I could not extract suitable paraphrases.

I obtain very good results for `AGENT` and `INSTRUMENT`, but other relations are problematic, probably due to the varying quality of the classifications: while *printer tray* and *film music* look correctly assigned to `CONTAINER`, *flood water* and *story idea* are quite abstract and questionable; *entrance stairs* (`DESTINATION`) could be equally well analyzed as `LOCATION` or `SOURCE`; and *exam anxiety* (`CAUSE`) could refer to `TIME`. Finally, although Table 4.4 shows the verb *to be* ranked third for *player coach*, in general the `EQUATIVE` relation poses a problem since the copula is not very frequent in the form of paraphrase I am looking for, e.g., '*coach who is a player*'.

4.6.3 Comparison to Rosario *et al.* (2002)

Rosario *et al.* (2002) characterize noun-noun compounds based on the semantic category in the MeSH lexical hierarchy each of the constituent nouns belongs to. For example, all noun compounds in which the first noun is classified under the A01 sub-hierarchy (*Body Regions*), and the second one falls under A07 (*Cardiovascular System*), are

[2]Barker & Szpakowicz (1998) allow for the modifier to be either a noun or an adjective.

Sem. Relation	Example	Extracted Verbs
POSSESSION	*family estate*	~~be in(29)~~, **be held by(9)**, **be owned by(7)**
TEMPORAL	*night flight*	**arrive at(19)**, **leave at(16)**, **be at(6)**, **be conducted at(6)**, **occur at(5)**
IS-A(HYPERNYMY)	*Dallas city*	**include(9)**
CAUSE	*malaria mosquito*	*carry(23)*, *spread(16)*, **cause(12)**, *transmit(9)*, *bring(7)*, *have(4)*, *be infected with(3)*, **be responsible for(3)**, *test positive for(3)*, **infect many with(3)**, *be needed for(3)*, **pass on(2)**, **give(2)**, **give out(2)**
MAKE/PRODUCE	*shoe factory*	**produce(28)**, **make(13)**, **manufacture(11)**
INSTRUMENT	*pump drainage*	**be controlled through(3)**, **use(2)**
LOCATION/SPACE	*Texas university*	~~be(5)~~, **be in(4)**
PURPOSE	*migraine drug*	**treat(11)**, **be used for(9)**, **prevent(7)**, **work for(6)**, **stop(4)**, **help(4)**, ~~work(4)~~ **be prescribed for(3)**, **relieve(3)**, **block(3)**, *be effective for(3)*, *be for(3)*, **help ward off(3)**, *seem effective against(3)*, **end(3)**, **reduce(2)**
SOURCE	*olive oil*	**come from(13)**, **be obtained from(11)**, **be extracted from(10)**, **be made from(9)**, **be produced from(7)**, **be released from(4)**, *taste like(4)*, **be beaten from(3)**, **be produced with(3)**, **emerge from(3)**
TOPIC	*art museum*	**focus on(29)**, *display(16)*, **bring(14)**, **highlight(11)**, *house(10)*, *exhibit(9)* **demonstrate(8)**, **feature(7)**, *show(5)*, **tell about(4)**, **cover(4)**, **concentrate in(4)**
MEANS	*bus service*	**use(14)**, **operate(6)**, *include(6)*
EXPERIENCER	*disease victim*	**spread(12)**, **acquire(12)**, **suffer from(8)**, **die of(7)**, *develop(7)*, **contract(6)**, **catch(6)**, **be diagnosed with(6)**, *have(5)*, *beat(5)*, **be infected by(4)**, **survive(4)**, **die from(4)**, **get(4)**, **pass(3)**, **fall by(3)**, *transmit(3)*
THEME	*car salesman*	**sell(38)**, ~~mean inside(13)~~, **buy(7)**, **travel by(5)**, **pay for(4)**, **deliver(3)**, **push(3)**, **demonstrate(3)**, **purr(3)**
RESULT	*combustion gas*	*support(22)*, **result from(14)**, **be produced during(11)**,**be produced by(8)** **be formed from(8)**, **form during(8)**, **be created during(7)**, **originate from(6)**, **be generated by(6)**, **develop with(6)**

Table 4.3: **Top paraphrasing verbs for 14 of the 21 relations in Girju *et al.* (2005).** Verbs expressing the target relation are in **bold**, those referring to a different but semantically valid one are in *italic*, and errors are ~~struck out~~.

Sem. Relation	Example	Extracted Verbs
AGENT	*student protest*	**be led by(6)**, **be sponsored by(6)**, **pit(4)**, *be(4)*, **be organized by(3)**, **be staged by(3)**, **be launched by(3)**, **be started by(3)**, **be supported by(3)**, *involve(3)*, *arise from(3)*
AGENT	*band concert*	*feature(17)*, *capture(10)*, *include(6)*, **be given**, **by(6)**, *play of(4)*, *involve(4)*, ~~be than(4)~~ **be organized by(3)**, **be by(3)**, *start with(3)*
AGENT	*military assault*	**be initiated by(4)**, *shatter(2)*
BENEFICIARY	*student price*	*be(14)*, ~~mean(4)~~, ~~differ from(4)~~, **be for(3)**, **be discounted for(3)**, **be affordable for(3)**, **be unfair for(3)**, **be charged for(3)**
CAUSE	*exam anxiety*	*be generated during(3)*
CONTAINER	*printer tray*	**hold(12)**, *come with(9)*, *be folded(8)*, *fit under(6)*, **be folded into(4)**, **pull from(4)**, **be inserted into(4)**, *be mounted on(4)*, *be used by(4)*, **be inside(3)**, **feed into(3)**
CONTAINER	*flood water*	*cause(24)*, *produce(9)*, *remain after(9)*, *be swept by(6)*, *create(5)*, *bring(5)*, *reinforce(5)*
CONTAINER	*film music*	*fit(16)*, **be in(13)**, **be used in(11)**, **be heard**, **in(11)**, *play throughout(9)*, *be written for(9)*
CONTAINER	*story idea*	*tell(20)*, *make(19)*, *drive(15)*, *become(13)*, *turn into(12)*, *underlie(12)*, **occur within(8)**, **hold(8)**, *tie(8)*, *be(8)*, *spark(8)*, *tell(7)*, *move(7)*
CONTENT	*paper tray*	**feed(6)**, *be lined with(6)*, *stand up(6)*, **hold(4)**, **contain(4)**, *catch(4)*, **overflow with(3)**
CONTENT	*eviction notice*	*result in(10)*, *precede(3)*, *make(2)*
DESTINATION	*game bus*	*be in(6)*, **leave for(3)**, *be like(3)*, *be(3)*, *make playing(3)*, *lose(3)*
DESTINATION	*exit route*	*be indicated by(4)*, **reach(2)**, *have(1)*, *do(1)*
DESTINATION	*entrance stairs*	*look like(4)*, *stand outside(3)*, *have(3)*, *follow from(3)*, *be at(3)*, ~~be(3)~~, *descend from(2)*
EQUATIVE	*player coach*	*work with(42)*, *recruit(28)*, **be(19)**, *have(16)*, *know(16)*, *help(12)*, *coach(11)*, *take(11)*
INSTRUMENT	*electron microscope*	**use(27)**, *show(5)*, **work with(4)**, **utilize(4)**, **employ(4)**, *beam(3)*
INSTRUMENT	*diesel engine*	*be(18)*, **operate on(8)**, *look like(8)*, **use(7)**, *sound like(6)*, **run on(5)**, **be on(5)**
INSTRUMENT	*laser printer*	**use(20)**, *consist of(6)*, *be(5)*

Table 4.4: **Comparison to Barker & Szpakowicz (1998): top paraphrasing verbs for 8 of the 20 relations.** Verbs expressing the target relation are in **bold**, those referring to a different but semantically valid one are in *italic*, and errors are ~~struck out~~.

Categ. Pair	Examples	Extracted Verbs
`A01-A07` `(Body Regions - Cardiovascular System)`	ankle artery foot vein forearm vein finger artery neck vein head vein leg artery thigh vein	*feed*(133), *supply*(111), *drain*(100), *be in*(44), *run*(37), *appear on*(29), *be located in*(22), *be found in*(20), *run through*(19), *be behind*(19), *run from*(18), *serve*(15), *be felt with*(14), *enter*(14), *pass through*(12), *pass by*(12), *show on*(11), *be visible on*(11), *run along*(11), *nourish*(10), *be seen on*(10), *occur on*(10), *occur in*(9), *emerge from*(9), *go into*(9), ...
`A01-M01.643` `(Body Regions - Disabled Persons)`	arm patient eye outpatient abdomen patient	*be*(54), *lose*(40), *have*(30), *be hit in*(11), *break*(9), *gouge out*(9), *injure*(8), *receive*(7), *be stabbed in*(7), *be shot in*(7), *need*(6), ...
`A01-M01.150` `(Body Regions - Disabled Persons)`	leg amputee arm amputee knee amputee	*lose*(13), *grow*(6), *have cut off*(4), *miss*(2), *need*(1), *receive*(1), *be born without*(1)
`A01-M01.898` `(Body Regions - Donors)`	eye donor skin donor	*give*(4), *provide*(3), *catch*(1)
`D02-E05.272` `(Organic Chemicals - Diet)`	choline diet methionine diet carotene diet saccharin diet	*be low in*(18), *contain*(13), *be deficient in*(11), *be high in*(7), *be rich in*(6), *be sufficient in*(6), *include*(4), *be supplemented with*(3), *be in*(3), *be enriched with*(3), *contribute*(2), *miss*(2), ...

Table 4.5: **Top paraphrasing verbs for relations from Rosario *et al.* (2002).**

hypothesized to express the same relation. Examples include *mesentery artery*, *leg vein*, *finger capillary*, etc.

Given a category pair of *MeSH* labels, I compare my approach to the descent of hierarchy by generating paraphrasing verbs on a large scale for many different compounds belonging to the target category pair.

I collect noun-noun compounds using the *LQL* system and the collection of 1.4 million *MEDLINE* abstracts described in chapter C in the appendix. I use the heuristic proposed by Lauer (1995), who extracts noun-noun pairs $w_1 w_2$ from four-word sequences $w_0 w_1 w_2 w_3$, where w_1 and w_2 are nouns and w_0, w_3 are non-nouns, with the additional requirement that both w_1 and w_2 represent single-word *MeSH* terms. The comparisons against *MeSH* are performed using all inflections and synonyms for a given term that are listed in *MeSH*. As a result, I obtain 228,702 noun-noun pairs, 40,861 of which are unique, which corresponds to 35,205 unique *MeSH* category pairs of various generalization levels.

Given a category pair, e.g., A01-A07, I consider all noun-noun compounds whose elements are in the corresponding *MeSH* sub-hierarchies, and I acquire paraphrasing verbs (+prepositions) for each of them from the Web. I then aggregate the results in order to obtain a set of characterizing paraphrasing verbs for the target category pair.

As Table 4.5 shows, the results are quite good for A01-A07, for which I have a lot of examples, and for D02-E05.272, which seems relatively unambiguous, but they are not as good for A01-M01.*, which is both more ambiguous and has fewer examples: generalizing verbal paraphrases for a category seems to work best for categories represented by multiple relatively unambiguous examples.

4.7 Comparison to Human-Generated Verbs

To evaluate the verb-based semantic relations I obtained, I conducted an experiment in which I gathered paraphrases from human judges. For the purpose, I defined a special noun-noun compound paraphrasing task asking human subjects to propose verbal paraphrases of the kind my program generates: I asked for verbs, possibly followed by prepositions, that could be used in a paraphrase involving *that*. For example, *nourish*, *run along* and *come from* are good paraphrasing verbs for the noun-noun compound *neck vein* since they can be used in paraphrases like '*a vein that* <u>*nourishes*</u> *the neck*', '*a vein that* <u>*runs along*</u> *the neck*' or '*a vein that* <u>*comes from*</u> *the neck*'. In an attempt to make the task as clear as possible and to ensure high quality of the results, I provided detailed instructions, I stated explicit restrictions, and I gave several example paraphrases. I instructed the participants to propose at least three paraphrasing verbs per noun-noun compound, if possible. The instructions I provided and the actual interface the human subjects were seeing are shown in Figures 4.1 and 4.2: worker's user interface of the *Amazon Mechanical Turk* Web service.[3]

The service represents a cheap and easy way to recruit subjects for various tasks that require human intelligence; it provides an API allowing a computer program to ask a human to perform a task and return the results, which *Amazon* calls "*artificial artificial intelligence*". The idea behind the latter term and behind the origin of service's name come from the "mechanical Turk", a life-sized wooden chess-playing mannequin the Hungarian nobleman Wolfgang von Kempelen constructed in 1769, which was able to defeat skilled

[3]`http://www.mturk.com`

Paraphrasing Noun-Noun Compounds

Introduction

Given a noun-noun compound like *malaria mosquito*, *olive oil*, *grain alcohol*, *canola leaves*, *fruit fly*, *evening ride*, *neck vein*, *disease victim*, *migraine drug*, *Google ads*, *etc.*, you are asked to paraphrase it using verbs and prepositions.
For example, *neck vein* can be paraphrased as follows:

"*neck vein*" is a *vein* that **comes from** the *neck*
"*neck vein*" is a *vein* that **drains** the *neck*
"*neck vein*" is a *vein* that **descends in** the *neck*
"*neck vein*" is a *vein* that **emerges from** the *neck*
"*neck vein*" is a *vein* that **enters** the *neck*
"*neck vein*" is a *vein* that **feeds** the *neck*
"*neck vein*" is a *vein* that **flows in** the *neck*
"*neck vein*" is a *vein* that **is in** the *neck*
"*neck vein*" is a *vein* that **is located in** the *neck*
"*neck vein*" is a *vein* that **is found in** the *neck*
"*neck vein*" is a *vein* that **is terminated at** the *neck*
"*neck vein*" is a *vein* that **nourishes** the *neck*
"*neck vein*" is a *vein* that **passes through** the *neck*
"*neck vein*" is a *vein* that **runs through** the *neck*
"*neck vein*" is a *vein* that **runs from** the *neck*
"*neck vein*" is a *vein* that **runs along** the *neck*
"*neck vein*" is a *vein* that **goes into** the *neck*
"*neck vein*" is a *vein* that **supplies** the *neck*
"*neck vein*" is a *vein* that **terminates in** the *neck*
etc.

Figure 4.1: **The noun-noun paraphrasing task in Amazon's Mechanical Turk:** task introduction.

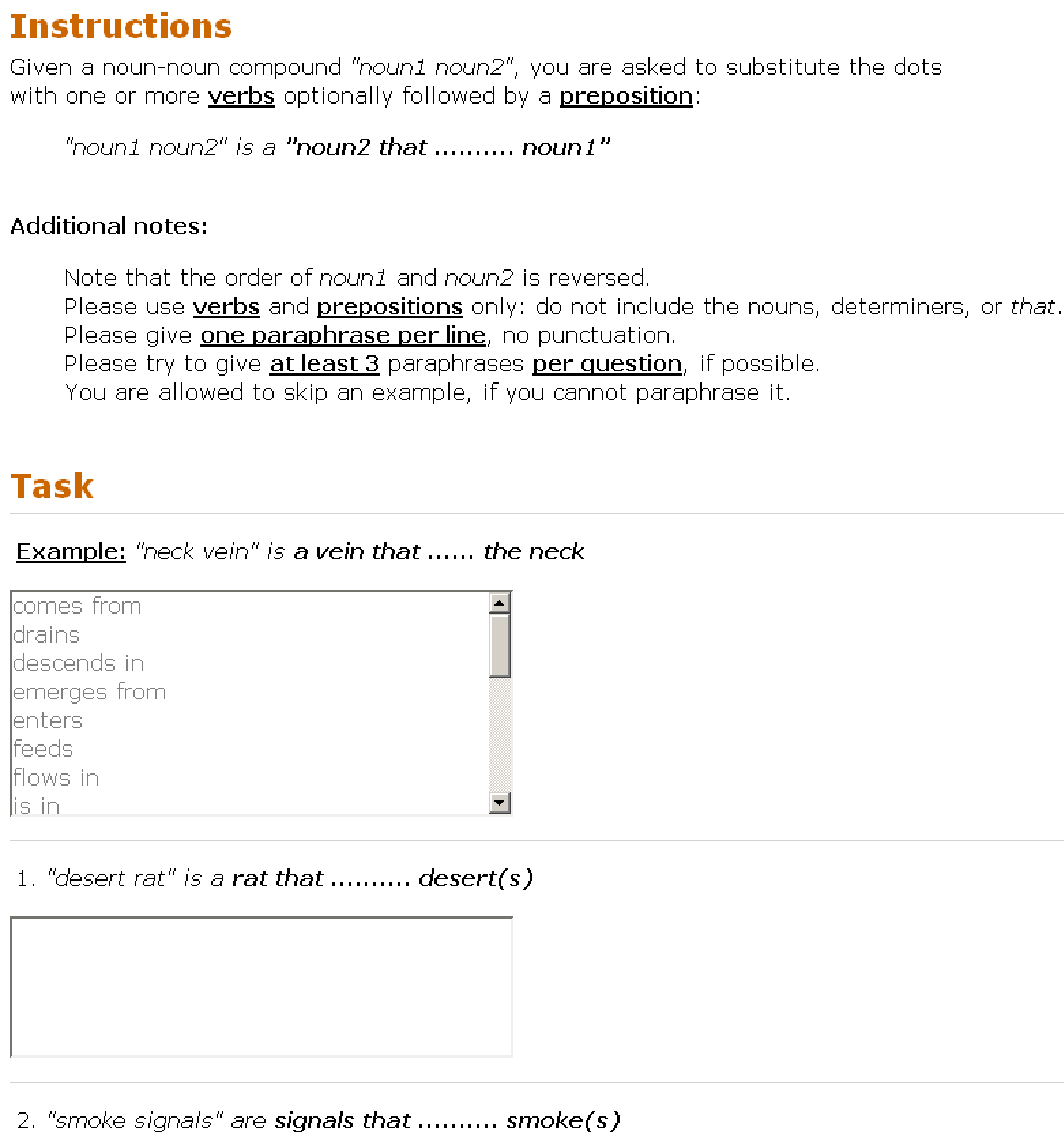

Figure 4.2: **The noun-noun paraphrasing task in Amazon's Mechanical Turk:** instructions, example, and sample questions.

opponents across Europe, including Benjamin Franklin and Napoleon Bonaparte. The audience believed the automaton was making decisions using artificial intelligence, but the actual secret was a chess master hidden inside. The *Amazon Mechanical Turk* Web Service provides a similar solution to computer applications.

Tables 4.6, 4.7 4.8 and 4.9 compare human- and program-generated paraphrasing verbs for *malaria mosquito*, *olive oil*, *disease victim* and *night flight*, respectively. The human-generated paraphrasing verbs, obtained from ten subjects, are shown on the left sides of the tables, while the right sides list the program-generated verbs; the verbs appearing on both sides are underlined. We can see in these tables a significant overlap between the human- and the program-generated paraphrasing verbs for *malaria mosquito*, and less overlap for the more ambiguous *night flight*, *disease victim*, and *olive oil*. For example, the latter can refer to multiple abstract relations, e.g., `CONTAINER` (*oil that <u>is inside</u> the olive*), `SOURCE` or `ORIGIN` (*oil that <u>comes from</u> olives*), `PRODUCT` (*oil that <u>is produced from</u> olives*), `QUALITY` (*oil that <u>tastes like</u> olive*), etc. Still, for all four given examples, there is a general tendency for the most frequent human-proposed and the top program-generated verbs to overlap.

I further compared the human and the program-generated paraphrases in a bigger study using the complex nominals listed in the appendix of Levi (1978). I had to exclude the examples with an adjectival modifier, which are allowed by Levi's theory, as I already explained in section 2.5.1. In addition, some of the noun compounds were spelled as a single word, which, according to my definition of noun compound given in section 2.4, represents a single noun. Therefore, I had to exclude the following concatenated words from

#	**Human Judges**	#	**Program**
8	carries	23	carries
4	causes	16	spreads
2	transmits	12	causes
2	is infected with	9	transmits
2	infects with	7	brings
1	has	4	has
1	gives	3	is infected with
1	spreads	3	infects with
1	propagates	2	gives
1	supplies	2	is needed for

Table 4.6: **Human- and program-generated verbs for *malaria mosquito*.** Verbs appearing on both sides of the tables are underlined.

#	**Human Judges**	#	**Program**
5	is pressed from	13	comes from
4	comes from	11	is obtained from
4	is made from	10	is extracted from
2	is squeezed from	9	is made from
2	is found in	7	is produced from
1	is extracted from	4	is released from
1	is in	4	tastes like
1	is produced out of	3	is beaten from
1	is derived from	3	is produced with
1	is created from	3	emerges from
1	contains		
1	is applied to		

Table 4.7: **Human- and program-generated verbs for *olive oil*.** Verbs appearing on both sides of the tables are underlined; both full and partial overlaps are underlined.

#	Human Judges	#	Program
6	has	12	spreads
3	suffers from	12	acquires
3	is infected with	8	suffers from
2	dies of	7	dies of
2	exhibits	7	develops
2	carries	6	contracts
1	is diagnosed with	6	is diagnosed with
1	contracts	6	catches
1	is inflicted with	5	has
1	is ill from	5	beats
1	succumbs to	4	is infected by
1	is affected by	4	survives
1	presents	4	dies from
		4	gets
		3	passes
		3	falls by
		3	transmits

Table 4.8: **Human- and program-generated verbs for *disease victim.*** Verbs appearing on both sides of the tables are underlined; both full and partial overlaps are underlined.

#	Human Judges	#	Program
5	occurs at	19	arrives at
5	is at	16	leaves at
4	happens at	6	is at
2	takes off at	6	is conducted at
1	arrives by	5	occurs at
1	travels through		
1	runs through		
1	occurs during		
1	is taken at		
1	is performed at		
1	is flown during		
1	departs at		
1	begins at		

Table 4.9: **Human- and program-generated verbs for *night flight.*** Verbs appearing on both sides of the tables are underlined; both full and partial overlaps are underlined.

Levi's dataset: *whistleberries*, *gunboat*, *silkworm*, *cellblock*, *snowball*, *meatballs*, *windmill*, *needlework*, *textbook*, *doghouse*, and *mothballs*. Some other examples contained a modifier that is a concatenated noun compound, e.g., *wastebasket category*, *hairpin turn*, *headache pills*, *basketball season*, *testtube baby*. These examples are noun-noun compounds under my definition, and therefore I retained them. However, I find them inconsistent with the other examples in the collection from Levi's theory point of view: the dataset is supposed to contain noun-noun compounds only. Even more problematic (but not for my definition), is *beehive hairdo*, where both the modifier and the head are concatenations; I retained that example as well. As a result, I ended up with 250 good noun-noun compounds out of the original 387 complex nominals.

I randomly distributed these 250 noun-noun compounds into groups of 5 as shown in Figure 4.2, which yielded 50 Mechanical Turk tasks known as HITs (Human Intelligence Tasks), and I requested 25 different human subjects (workers) per HIT. I had to reject some of the submissions, which were empty or were not following the instructions, in which cases I requested additional workers in order to guarantee at least 25 good submissions per HIT. Each human subject was allowed to work on any number of HITs (between 1 and 50), but was not permitted to do the same HIT twice, which is controlled by the *Amazon Mechanical Turk* Web Service. A total of 174 different human subjects worked on the 50 HITs, producing 19,018 different verbs. After removing the empty and the bad submissions, and after normalizing the verbs (see below), I ended up with a total of 17,821 verbs, which means 71.28 verbs per noun-noun compound on average, not necessarily distinct.

Since many workers did not strictly follow the instructions, I performed some automatic cleaning of the results, followed by a manual check and correction, when it was necessary. First, some workers included the target nouns, the complementizer *that*, or determiners like *a* and *the*, in addition to the paraphrasing verb, in which cases I removed this extra material. For example, *star shape* was paraphrased as *shape that looks like a star* or as *looks like a* instead of just *looks like*. Second, the instructions required that a paraphrase be a sequence of one or more verb forms possibly followed by a preposition (complex prepositions like *because of* were allowed), but in many cases the proposed paraphrases contained words belonging to other parts of speech, e.g., nouns (*is in the shape of*, *has responsibilities of*, *has the role of*, *makes people have*, *is part of*, *makes use of*) or predicative adjectives (*are local to*, *is full of*); I filtered out such paraphrases. In case a paraphrase contained an adverb, e.g., *occur only in*, *will eventually bring*, I removed the adverb and kept the paraphrase. Third, I normalized the verbal paraphrases by removing the leading modals (e.g., *can cause* becomes *cause*), perfect tense *have* and *had* (e.g., *have joined* becomes *joined*), or continuous tense *be* (e.g., *is donating* becomes *donates*). I converted complex verbal construction of the form '*<raising verb> to be*' (e.g., *appear to be*, *seems to be*, *turns to be*, *happens to be*, *is expected to be*) to just *be*. I further removed present participles introduced by *by*, e.g., *are caused by peeling* becomes *are caused*. I further filtered out any paraphrase that involved *to* as part of the infinitive of a verb different from *be*, e.g., *is willing to donate* or *is painted to appear like* are not allowed. I also added *be* when it was missing in passive constructions, e.g., *made from* became *be made from*. Finally, I lemmatized the conjugated verb forms using *WordNet*, e.g., *comes from* becomes *come from*, and *is produced from* becomes *be produced from*. I also fixed some occasional spelling errors that I noticed, e.g., *bolongs to*,

Min # of Web Verbs	Number of Compounds	Correlation to Human All Verbs	First Only
0	250	31.81%	30.60%
1	236	33.70%	32.41%
3	216	35.39%	34.07%
5	203	36.85%	35.60%
10	175	37.31%	35.53%

Table 4.10: **Average cosine correlation (in %s) between the human- and the program-generated verbs for 250 noun-noun compounds from Levi *et al.* (1978).** Shown are the results for different limits on the minimum number of program-generated Web verbs. The last column shows the cosine when only the first verb proposed by each worker is used.

happens becasue of, is mmade from.

For each noun-noun compound, I built two frequency vectors $\overrightarrow{h}$ and $\overrightarrow{p}$, using the human-generated paraphrasing verbs and their frequencies, and the program-generated ones, respectively. Each coordinate corresponds to I then calculated the cosine correlation coefficient between these frequency vectors as follows:

$$sim_{cos}(\overrightarrow{h}, \overrightarrow{p}) = \frac{\sum_{i=1}^{n} h_i p_i}{\sqrt{\sum_{i=1}^{n} h_i^2}\sqrt{\sum_{i=1}^{n} p_i^2}} \tag{4.1}$$

The average cosine correlation (in %s) for all 250 noun-noun compounds is shown in Table 4.10. Since the human subjects were instructed to provide at least three paraphrasing verbs per compound and they tried to comply, this sometimes yielded bad verbs. In such cases, the very first verb proposed by a worker for a given noun-noun compound is likely to be the best one. I tested this hypothesis by calculating the cosine using these first verbs only. As the last two columns of the table show, using all verbs produces a consistently better cosine, which suggests that there are many additional good human-generated verbs among the ones after the first. A quick comparison of sections E.1 and E.2 in the Appendix confirms this. However, the difference is small, about 1-2%.

A limitation of the Web-based verb-generating method is that it could not provide paraphrasing verbs for 14 of the noun-noun compounds, in which cases the cosine was zero. If the calculation was performed for the remaining 236 compounds only, the cosine increased by 2%. Table 4.10 shows the results when the cosine calculations are limited to compounds with at least 1, 3, 5 or 10 different verbs. We can see that the correlation increases with the minimum number of required verbs, which means that the extracted verbs are generally good, and part of the low cosines are due to an insufficient number of extracted verbs.

Overall, all cosines in Table 4.10 are in the 30-37%, which corresponds to a medium correlation (Cohen 1988). A detailed comparison for all 250 is shown in Appendix E: see section E.1 for the results when all human-proposed verbs are used, and section E.2 for only the first verb proposed by each worker is used.

I further compare the human- and the program-generated verbs aggregated by relation. Given a relation, e.g., $\texttt{HAVE}_1$, I collect all verbs belonging to noun-noun compounds from that relation together with their frequencies. From a vector-space model point of view, I sum their corresponding frequency vectors. I do that separately for the human- and the program-generated verbs, and I then compare them separately for each relation. A detailed comparison for all 16 relations is shown in Appendix E: see section E.3 for the results when all human-proposed verbs are used, and section E.4 for when only the first verb proposed by each worker is used.

Figure 4.3 shows the cosine correlations for each of the 16 relations using all human-proposed verbs and the first verb from each worker. We can see a very-high correlation (mid

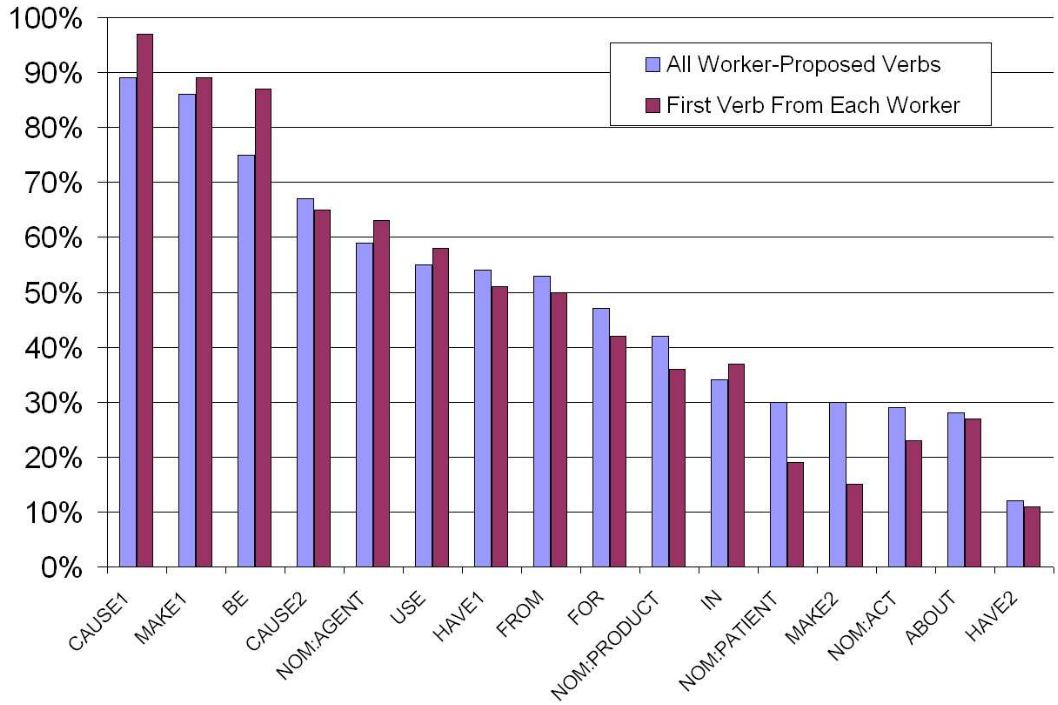

Figure 4.3: **Cosine correlation (in %s) between the human- and the program-generated verbs aggregated by relation:** using all human-proposed verbs vs. the first verb from each worker.

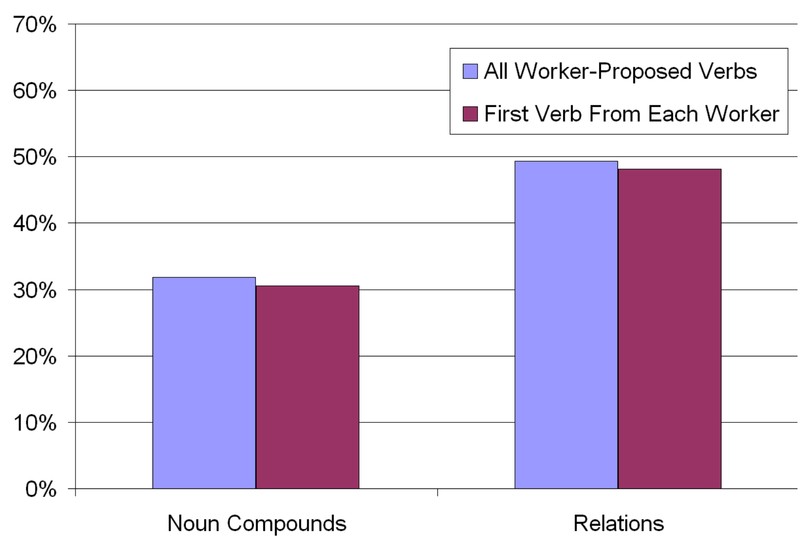

Figure 4.4: **Average cosine correlation (in %s) between the human- and the program-generated verbs for 250 noun-noun compounds from Levi *et al.* (1978) calculated for each noun compound and aggregated by relation:** using all human-proposed verbs vs. the first verb from each worker.

70s to mid 90s) for relations like $\texttt{CAUSE}_1$, $\texttt{MAKE}_1$, `BE`, but low correlation 11-30% for reverse relations like $\texttt{HAVE}_2$ and $\texttt{MAKE}_2$, for rare relations like `ABOUT`, and for most nominalizations (except for `NOM:AGENT`). Interestingly, using the first verb only improves the results for the highly-correlated relations, but damages the low-correlated ones. This suggests that when a relation is more homogeneous, the first verbs proposed by the workers are good enough, and the following verbs only introduce noise. However, when the relation is more heterogeneous, the extra verbs become more likely to be useful. As Figure 4.4 shows, overall the average cosine correlation is slightly higher when all worker-proposed verbs are used vs. the first verb from each worker only: this is true both when comparing the individual noun-noun compounds and when the comparison is performed for the 16 relations. The figure also shows that while the cosine correlation for the individual noun-noun compounds is in the low 30s, for the relations it is almost 50%.

4.8 Comparison to FrameNet

The idea to approximate noun compounds' semantics by a collection of verbs is related to the approach taken in the *Berkeley FrameNet project*[4] (Baker *et al.* 1998), which builds on the ideas for case frames of Fillmore (1968). According to Fillmore (1982), frames are "*characterizing a small abstract 'scene' or 'situation', so that to understand the semantic structure of the verb it is necessary to understand the properties of such schematized scenes*". In *FrameNet*, a lexical item evokes a number of other relations and concepts representative of the context in which the word applies; together these make up the speaker's understanding of

[4] http://www.icsi.berkeley.edu/∼framenet/

Frames	**Verbs**	**Program**	**Humans**
Causation	(be) because of		✓
Causation	bring	✓	+
Causation	bring about	✓	✓
Causation	bring on	+	✓
Causation	induce	✓	✓
Causation	lead to	+	✓
Causation	make	✓	✓
Causation	mean		
Causation	precipitate		✓
Causation	put		
Causation	raise		
Causation	result in	✓	✓
Causation	render		
Causation	send		
Causation	wreak		
Overlap:		**5/15 (33.3%)**	**8/15 (53.3%)**

Table 4.11: ***Causation* frame and $\texttt{CAUSE}_1$:** comparing *FrameNet* with the top 150 program- and the human-generated verbs for $\texttt{CAUSE}_1$.

Frame	**Verb**	**Program**	**Humans**
Using	apply	✓	✓
Using	avail oneself		
Using	employ	✓	✓
Using	operate	✓	+
Using	utilise	✓	✓
Using	use	✓	✓
Overlap:		**5/6 (83.3%)**	**4/6 (66.7%)**

Table 4.12: ***Using* frame and `USE`:** comparing *FrameNet* with the top 150 program- and the human-generated verbs for `USE`.

Frames	**Verbs**	**Program**	**Humans**
Possession	belong		+
Possession, Have_associated	have got		
Possession, Have_associated	have	✓	✓
Possession	lack	✓	
Possession	own	✓	✓
Possession	possess	✓	
Possession	want	✓	
Inclusion	(be) among		
Inclusion	(be) amongst		
Inclusion	contain	✓	✓
Inclusion	exclude		
Inclusion	(be) excluding		
Inclusion	incorporate	✓	✓
Overlap:		**7/13 (53.8%)**	**4/13 (30.8%)**

Table 4.13: ***Possession*, *Inclusion* and *Have_associated* frames and** $\texttt{HAVE}_1$**:** comparing *FrameNet* with the top 150 program- and the human-generated verbs for $\texttt{HAVE}_1$.

Frames	**Verbs**	**Program**	**Humans**
Intentionally_create	create	✓	✓
Intentionally_create	establish		
Intentionally_create	found		
Intentionally_create	generate	✓	✓
Intentionally_create, Manufacturing	make	✓	✓
Intentionally_create, Manufacturing	produce	✓	✓
Intentionally_create	setup		
Intentionally_create	synthesise		✓
Manufacturing	fabricate		
Manufacturing	manufacture		✓
Overlap:		**4/10 (40.0%)**	**6/10 (60.0%)**

Table 4.14: ***Intentionally_create* and *Manufacturing* frames and** $\texttt{MAKE}_1$**:** comparing *FrameNet* with the top 150 program- and the human-generated verbs for $\texttt{MAKE}_1$.

the word. For example, the meaning of a sentence such as '*Sara faxed Jeremy the invoice.*' is not derived from the meaning of the verb *fax* alone, but also from speakers' knowledge about situations where somebody gives something to somebody else (Goldberg 1995; Petruck 1996; Baker & Ruppenhofer 2002).

In a sense, the above-described program- and human-generated paraphrasing verbs represent exactly such kind of world knowledge: a dynamically constructed semantic frame in terms of which the target noun compound is to be understood. Therefore, it would not be unreasonable to expect similarity of the relations between the entities in the manually created frames of *FrameNet* and the ones I generate dynamically from the Web: in particular, there should be some overlap between my automatically generated verbs and the ones listed in the corresponding *FrameNet* frame. If so, given a sentence, my verbs can help automatically select the *FrameNet* frame that best characterizes the situation described in that sentence.

As a preliminary investigation of the potential of these ideas, I compared the verbs I generate for four of Levi's relations $\texttt{CAUSE}_1$, $\texttt{USE}$, $\texttt{HAVE}_1$, and $\texttt{MAKE}_1$, and the verbs listed in *FrameNet* for the frames, I found to correspond to these relations. I also tried to compare the *FrameNet* verbs to the human-proposed ones. In both cases, I used the top 150 verbs for the target Levi relation, as shown in appendix E.3. The results are shown in Tables 4.11, 4.12, 4.13, and 4.14, respectively. The results vary per relation, but overall the performance of the human- and program-proposed verbs is comparable, and the average overlap is over 50%, which is quite promising.

4.9 Application to Relational Similarity

In this section, I extend the above method which characterizes the semantic relations that hold between nouns in noun-noun compounds, to measuring the semantic similarity between pairs of words, i.e. relational similarity. The approach leverages the vast size of the Web to building linguistically-motivated lexical features. Given a pair of words, it mines the Web for sentences containing these words and then extracts verbs, prepositions, and coordinating conjunctions that connect them. These lexical features are then used in a vector-space model to measure semantic similarity that is needed for building instance-based classifiers. The evaluation of the approach on several relational similarity problems, including SAT verbal analogy, head-modifier relations, and relations between complex nominals shows state-of-the-art performance.

4.9.1 Introduction

Measuring the semantic similarity between pairs of words, i.e. relational similarity, is an important but understudied problem. Despite the tremendous amount of computational linguistics publications on word similarity (see (Budanitsky & Hirst 2006) for an overview), there is surprisingly few work on relational similarity. Students taking the SAT examination are familiar with verbal analogy questions, where they need to decide whether, e.g., the relation between *ostrich* and *bird* is more similar to the one between *lion* and *cat*, or rather between *primate* and *monkey*. These kinds of questions are hard; the average test taker achieves about 56.8% on the average (Turney & Littman 2005).

Many NLP applications would benefit from solving the relational similarity problem, e.g., question answering, information retrieval, machine translation, word sense disambiguation, information extraction, etc. While there are few success stories so far, the ability to measure semantic similarity has already proven its advantages for textual entailment (Tatu & Moldovan 2005), and the importance of the task is being realized: there was a competition in 2007 on *Classification of Semantic Relations between Nominals* as part of *SemEval*, and the journal of *Language Resources and Evaluation* will have a special issue in 2008 on *Computational Semantic Analysis of Language: SemEval-2007 and Beyond*.

Below I introduce a novel Web-based approach, which, despite its simplicity, rivals the best previous approaches to relational similarity. Following Turney (2006b), I use SAT verbal analogy as a benchmark problem. I further address two semantic relation classification problems: head-modifier relations, and relations between complex nominals.

4.9.2 Method

Feature Extraction

Given a pair of nouns $noun_1$ and $noun_2$, I mine the Web for sentences containing them and then I extract connecting verbs, prepositions, and coordinating conjunctions, which I will later use as lexical features in a vector-space model to measure semantic similarity between pairs of nouns.

The extraction process starts with a set of exact phrase queries generated using the following patterns:

$$\text{“}infl_1 \texttt{ THAT } * \ infl_2\text{”}$$

$$\text{“}infl_2 \texttt{ THAT } * \ infl_1\text{”}$$

$$\text{“}infl_1 \ * \ infl_2\text{”}$$

$$\text{“}infl_2 \ * \ infl_1\text{”}$$

where:

$infl_1$ and $infl_2$ are inflected variants of $noun_1$ and $noun_2$;

`THAT` can be *that*, *which*, or *who*;

and `*` stands for 0 or more (up to 8) stars, representing the *Google* `*` operator.

For each query, I collect the text snippets (summaries) from the result set (up to 1,000 per query) and I split them into sentences. I then filter out the incomplete sentences and the ones that do not contain the target nouns, I assign POS annotations using the OpenNLP tagger[5], and I extract the following features:

Verb: I extract a verb, if the subject NP of that verb is headed by one of the target nouns (or an inflected form of a target noun), and its direct object NP is headed by the other target noun (or an inflected form). For example, the verb *include* will be extracted from "The *committee* <u>includes</u> many *members*." I also extract verbs from relative clauses, e.g., "This is a *committee* which <u>includes</u> many *members*." Verb particles are also recognized, e.g., "The *committee* must <u>rotate off</u> 1/3 of its *members*." I ignore modals and auxiliaries, but retain the passive *be*. Finally, I lemmatize the main verb using *WordNet*'s morphological analyzer *Morphy* (Fellbaum 1998).

Verb+Preposition: If the subject NP of a verb is headed by one of the target nouns (or an inflected form), and its indirect object is a PP containing an NP which is headed by the other target noun (or an inflected form), I extract the verb and the preposition

[5]OpenNLP: `http://opennlp.sourceforge.net`

heading that PP, e.g., "The thesis advisory *committee* consists of three qualified *members*." As in the verb case, I also extract verb+preposition from relative phrases, I include particles, I ignore modals and auxiliaries, and I lemmatize the verbs.

Preposition: If one of the target nouns is the head of an NP that contains a PP inside which there is an NP headed by the other target noun (or an inflected form), I extract the preposition heading that PP, e.g., "The *members* of the *committee* held a meeting."

Coordinating conjunction: If the two target nouns are the heads of coordinated NPs, I extract the coordinating conjunction.

In addition to the lexical part, for each extracted feature, I keep a direction. Therefore the preposition *of* represents two different features in the following examples "*member* of the *committee*" and "*committee* of *members*". See Table 4.15 for examples.

The proposed method is similar in spirit to previous paraphrase acquisition approaches which search for similar/fixed endpoints and collect the intervening material. Lin & Pantel (2001) use a dependency parser and extract paraphrases from dependency tree paths whose ends contain semantically similar sets of words by generalizing over these ends. For example, given the target phrase "*X solves Y*", they extract paraphrases such as "*X finds a solution to Y*", "*X tries to solve Y*", "*X resolves Y*", "*Y is resolved by X*", etc. The approach is extended by Shinyama *et al.* (2002), who use named entity recognizers and look for anchors belonging to a matching semantic class, e.g., `LOCATION`, `ORGANIZATION`, etc. This latter idea is further extended by Nakov *et al.* (2004), who apply it in the biomedical domain, imposing the additional restriction that the sentences from which the paraphrases are to be extracted represent citation sentences (citances) that cite the same target paper;

Frequency	**Feature**	**POS**	**Direction**
2205	of	P	$2 \rightarrow 1$
1923	be	V	$1 \rightarrow 2$
771	include	V	$1 \rightarrow 2$
382	serve on	V	$2 \rightarrow 1$
189	chair	V	$2 \rightarrow 1$
189	have	V	$1 \rightarrow 2$
169	consist of	V	$1 \rightarrow 2$
148	comprise	V	$1 \rightarrow 2$
106	sit on	V	$2 \rightarrow 1$
81	be chaired by	V	$1 \rightarrow 2$
78	appoint	V	$1 \rightarrow 2$
77	on	P	$2 \rightarrow 1$
66	and	C	$1 \rightarrow 2$
66	be elected	V	$1 \rightarrow 2$
58	replace	V	$1 \rightarrow 2$
48	lead	V	$2 \rightarrow 1$
47	be intended for	V	$1 \rightarrow 2$
45	join	V	$2 \rightarrow 1$
45	rotate off	V	$2 \rightarrow 1$
44	be signed up for	V	$2 \rightarrow 1$
43	notify	V	$1 \rightarrow 2$
40	provide that	V	$2 \rightarrow 1$
39	need	V	$1 \rightarrow 2$
37	stand	V	$2 \rightarrow 1$
36	be	V	$2 \rightarrow 1$
36	vote	V	$1 \rightarrow 2$
36	participate in	V	$2 \rightarrow 1$
35	allow	V	$1 \rightarrow 2$
33	advise	V	$2 \rightarrow 1$
32	inform	V	$1 \rightarrow 2$
31	form	V	$2 \rightarrow 1$
...	...	...	...

Table 4.15: **The most frequent Web-derived features for *committee member*.** Here V stands for verb (possibly +preposition and/or +particle), P for preposition and C for coordinating conjunction; $1 \rightarrow 2$ means *committee* precedes the feature and *member* follows it; $2 \rightarrow 1$ means *member* precedes the feature and *committee* follows it.

this restriction yields higher accuracy. While the objective of all these approaches is paraphrase acquisition, here I am interested in extracting linguistic features that can express the semantic relation between two nouns.

After having extracted the linguistic features, I use them in a vector-space model in order to measure semantic similarity between pairs of nouns. This vector representation is similar to the ones proposed by Alshawi & Carter (1994), Grishman & Sterling (1994), Ruge (1992), and Lin (1998). For example, the latter measures *word* similarity using triples extracted from a dependency parser. In particular, given a noun, Lin (1998) finds all verbs that have it as a subject or object, and all adjectives that modify it, together with the corresponding frequencies. Unlike this research, whose objective is measuring word similarity, here I am interested in relational similarity.

The method is also similar to the ideas of Devereux & Costello (2006) that the meaning of a noun-noun compound is characterized by a distribution over several dimensions, as opposed to be expressed by a single relation. However, their relations are fixed and abstract, while mine are dynamic and based on verbs, prepositions and coordinating conjunctions. It is also similar to Kim & Baldwin (2006), who characterize the relation using verbs; however they use a fixed set of seed verbs, and my verbs are dynamically extracted. My approach is also similar to that of Séaghdha & Copestake (2007) who use grammatical relations as features to characterize a noun-noun compound; however I use verbs, prepositions and coordinating conjunctions instead.

Similarity Measure

The above-described features are used in the calculation of the similarity between noun pairs. I use TF.IDF-weighting in order to downweight very common features like *of*:

$$w(x) = TF(x) \times \log\left(\frac{N}{DF(x)}\right) \tag{4.2}$$

In the above formula, $TF(x)$ is the number of times the feature x has been extracted for the target noun pair, $DF(x)$ is the total number of training noun pairs that have this feature, and N is the total number of training noun pairs.

I use these weights in a variant of the Dice coefficient. The classic Dice coefficient for two sets A and B is defined as follows:

$$Dice(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{4.3}$$

This definition applies to Boolean vectors as well since they are equivalent to discrete sets, but it does not apply to numerical vectors in general. Therefore, I use the following generalized definition[6]:

$$Dice(A, B) = \frac{2 \times \sum_{i=1}^{n} \min(a_i, b_i)}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} \tag{4.4}$$

A bigger value for the Dice coefficient indicates higher similarity. Therefore, I take $\min(a_i, b_i)$ in order to avoid giving unbalanced weight to a feature when the weights are lopsided. For example, if the noun pair (*committee*, *members*) has the feature *include* as a verb in the forward direction 1,000 times, and the noun pair (*ant*, *hill*) has it only twice, I do not want to give a lot of weight for overlapping on that feature.

[6] Other researchers have proposed different extensions, e.g., Lin (1998).

	ostrich:bird		**palatable:toothsome**
(a)	*lion:cat*	(a)	rancid:fragrant
(b)	goose:flock	(b)	chewy:textured
(c)	ewe:sheep	*(c)*	*coarse:rough*
(d)	cub:bear	(d)	solitude:company
(e)	primate:monkey	(e)	no choice

Table 4.16: **SAT verbal analogy: sample questions.** The stem is in **bold**, the correct answer is in *italic*, and the distractors are in plain text.

4.9.3 SAT Verbal Analogy

Following Turney (2006b), I use *SAT verbal analogy* as a benchmark problem. I experiment with *Turney's dataset*, which consists of 374 SAT questions from various sources, including 190 from actual SAT tests, 80 from SAT guidebooks, 14 from the ETS Web site, and 90 from other SAT test preparation Web sites. Table 4.16 shows two example problems: the top pairs are called *stems*, the ones in italic are the *solutions*, and the remaining ones are *distractors*. Turney (2006b) achieves 56% accuracy, which matches the 56.8% average human performance, and is a significant improvement over the 20% random-guessing baseline.

Note that the righthand example in Table 4.16 misses one distractor; so do 21 examples in *Turney's dataset*. It also mixes different parts of speech: while *solitude* and *company* are nouns, all remaining words are adjectives. Other examples in *Turney's dataset* contain verbs and adverbs, and even relate pairs of different part of speech. This is problematic for my approach, which requires that both words be nouns[7]. After having filtered all examples containing non-nouns, I ended up with 184 questions, which I use in the evaluation.

[7]The approach can be extended to handle adjective-noun pairs, as demonstrated in section 4.9.4 below.

Model	Correct	Wrong	N/A	Accuracy	Coverage
$v+p+c$	129	52	3	**71.27±6.98**	**98.37**
v	122	56	6	68.54±7.15	96.74
$v+p$	119	61	4	66.11±7.19	97.83
$v+c$	117	62	5	65.36±7.23	97.28
$p+c$	90	90	4	50.00±7.23	97.83
p	84	94	6	47.19±7.20	96.74
baseline	37	147	0	20.00±5.15	100.00
LRA: (Turney 2006b)	122	59	3	67.40±7.13	98.37

Table 4.17: **SAT verbal analogy: evaluation on 184 noun-only questions.** For each model, the number of correctly classified, wrongly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). The letters v, p and c indicate the kinds of features used: v stands for verb (possibly +preposition), p for preposition, and c for coordinating conjunction.

Given a SAT verbal analogy example, I build six feature vectors – one for each of the six word pairs. I calculate the similarity between the stem of the analogy and each of the five candidates using the Dice coefficient with TF.IDF-weighting, and I choose the pair with the highest score; I make no prediction if two or more pairs tie for the highest score.

The evaluation results are shown in Table 4.17: I use leave-one-out cross validation since I need to set the TF.IDF weights. The last line shows the performance of Turney's Latent Relational Analysis (LRA) when limited to the 184 noun-only dataset. My best model $v+p+c$ performs a bit better, 71.27% vs. 67.40%, but the difference is not statistically significant (Pearson's Chi-square test, see section 3.9.5). Note that this "inferred" accuracy could be misleading, and the LRA would have performed better if it was trained to solve *noun-only* analogies, which seem easier, as demonstrated by the significant increase in accuracy for LRA when limited to nouns: 67.4% vs. 56.8% for 184 and 374 questions, respectively. The learning process for LRA is probably harder on the full dataset, e.g., its pattern discovery step needs to learn patterns for many POS as opposed to nouns only, etc.

Model	Correct	Wrong	N/A	Accuracy	Coverage
$v + p$	240	352	8	**40.54±3.88**	**98.67**
$v + p + c$	238	354	8	40.20±3.87	98.67
v	234	350	16	40.07±3.90	97.33
$v + c$	230	362	8	38.85±3.84	98.67
$p + c$	114	471	15	19.49±3.01	97.50
p	110	475	15	19.13±2.98	97.50
baseline	49	551	0	8.17±1.93	100.00
LRA (Turney)	239	361	0	39.83±3.84	100.00

Table 4.18: **Head-modifier relations, 30 classes: evaluation on *Barker&Szpakowicz dataset.*** For each model, the number of correctly classified, wrongly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). Accuracy and coverage are micro-averaged.

4.9.4 Head-Modifier Relations

Next, I experiment with a head-modifier dataset by Barker & Szpakowicz (1998), which contains head-modifier relations between noun-noun, adjective-noun and adverb-noun pairs. The dataset contains 600 head-modifier examples, each annotated with 30 fine-grained relations, grouped into 5 coarse-grained classes (the covered fine-grained relations are shown in parentheses): CAUSALITY (*cause, effect, purpose, detraction*), TEMPORALITY (*frequency, time_at, time_through*), SPATIAL (*direction, location, location_at, location_from*), PARTICIPANT (*agent, beneficiary, instrument, object, object_property, part, possessor, property, product, source, stative, whole*) and QUALITY (*container, content, equative, material, measure, topic, type*). For example, *exam anxiety* is classified as *effect* and therefore as CASUALITY, and *blue book* is *property* and therefore also PARTICIPANT.

There are some problematic examples in this dataset. First, in three cases, there are two modifiers rather than one, e.g., *infectious disease agent*. In these cases, I ignore the first modifier. Second, seven examples have an adverb as a modifier, e.g., *daily exercise*,

Model	Correct	Wrong	N/A	Accuracy	Coverage
$v + p$	328	264	8	**55.41±4.03**	**98.67**
$v + p + c$	324	269	7	54.64±4.02	98.83
v	317	267	16	54.28±4.06	97.33
$v + c$	310	280	10	52.54±4.03	98.33
$p + c$	240	345	15	41.03±3.91	97.50
p	237	341	22	41.00±3.94	96.33
baseline	260	340	0	43.33±3.91	100.00
LRA (Turney)	348	252	0	58.00±3.99	100.00

Table 4.19: **Head-modifier relations, 5 classes: evaluation on *Barker&Szpakowicz dataset.*** For each model, the number of correctly classified, wrongly classified, and non-classified examples is shown, followed by the accuracy (in %) and the coverage (% of examples for which the model makes prediction). Accuracy and coverage are micro-averaged.

and 262 examples have an adjective as a modifier, e.g., *tiny cloud*. I treat them as if the modifier was a noun, which works in many cases, since many adjectives and adverbs can be used predicatively, e.g., '*This exercise is performed daily.*' or '*This cloud looks very tiny.*'

For the evaluation, I create a feature vector for each head-modifier pair, and I perform a leave-one-out cross-validation: I leave one example for testing and I train on the remaining 599; I repeat this procedure 600 times, so that each example gets used for testing. Following Turney & Littman (2005) and Barker & Szpakowicz (1998), I use a 1-nearest-neighbor classifier. I calculate the similarity between the feature vector of the testing example and the vectors of the training examples. If there is a single highest-scoring training example, I predict its class for that test example. Otherwise, if there are ties for first, I assume the class predicted by the majority of the tied examples, if there is a majority.

The results for the 30-class and 5-class *Barker&Szpakowicz dataset* are shown in Tables 4.18 and 4.19, respectively. For the 30-way classification, my best model achieves 40.54% accuracy, which is comparable to the accuracy of Turney's LRA: 39.83%. For the

5-way classification, I achieve 55.41% vs. 58.00% for Turney's LRA. In either case, the differences are not statistically significant (tested with Pearson's Chi-square test, see section 3.9.5). Given that Turney's algorithm requires substantial resources, synonym expansion, and costly computation over multiple machines, I believe that my simple approach is preferable.

Overall, for *Barker&Szpakowicz dataset*, it is best to use verbs and prepositions ($v+p$); adding coordinating conjunctions ($v+p+c$) lowers the accuracy. Using prepositions in addition to verbs ($v + p$) is better than using verbs only (v), but combining verbs and coordinating conjunctions ($v + c$) lowers the accuracy. Coordinating conjunctions only help when combined with prepositions ($p + c$). Overall, verbs are the most important features, followed by prepositions.

The reason coordinating conjunctions cannot help increase the accuracy is that head-modifier relations are typically expressed with verbal or prepositional paraphrase; coordinating conjunctions only help with some infrequent relations, like *equative*, e.g., finding *player and coach* suggests an equative relation for *player coach* or *coach player*.

This is different for SAT verbal analogy, where the best model is $v + p + c$, as Table 4.17 shows. Verbs are still the most important feature and also the only one whose presence/absence makes a statistical difference (as the confidence intervals show). However, this time using c does help: SAT verbal analogy questions ask for a broader range of relations, e.g., antonyms, for which coordinating conjunctions like *but* can be helpful.

#	Relation Name	Examples	Train	Test
1	Cause-Effect	hormone-growth, laugh-wrinkle	140	80
2	Instrument-Agency	laser-printer, ax-murderer	140	78
3	Product-Producer	honey-bee, philosopher-theory	140	93
4	Origin-Entity	grain-alcohol, desert-storm	140	81
5	Theme-Tool	work-force, copyright-law	140	71
6	Part-Whole	leg-table, door-car	140	72
7	Content-Container	apple-basket, plane-cargo	140	74

Table 4.20: **SemEval dataset**: relations with examples (context sentences are not shown). Also shown are the number of training and testing instances for each relation.

4.9.5 Relations Between Nominals

The last dataset I experiment with is from *SemEval'2007* competition Task #4: *Classification of Semantic Relations between Nominals* (Girju *et al.* 2007). It contains a total of seven semantic relations, not exhaustive and possibly overlapping, with 140 training and about 70 testing examples per relation. Table 4.20 lists the seven relations. This is a binary classification task and each relation is considered in isolation; there are approximately 50% negative and 50% positive examples ("near misses") per relation.

Each example consists of a sentence, the target semantic relation, two nominals to be judged on whether they are in that relation, manually annotated WordNet 3.0 sense keys for these nominals, and the Web query used to obtain that example:

```
"Among the contents of the <e1>vessel</e1>
were a set of carpenters <e2>tools</e2>,
several large storage jars, ceramic
utensils, ropes and remnants of food, as
well as a heavy load of ballast stones."
WordNet(e1) = "vessel%1:06:00::",
WordNet(e2) = "tool%1:06:00::",
Content-Container(e2, e1) = "true",
Query = "contents of the * were a"
```

Given an entity-annotated example sentence, I reduce the target entities e_1 and e_2 to single nouns $noun_1$ and $noun_2$ keeping their last nouns only, which I assume to be the heads. I then mine the Web for sentences containing both $noun_1$ and $noun_2$, and I build feature vectors as above. In addition to the Web-derived features, I use the following contextual ones:

Sentence word: I use as features the words from the context sentence, after stop words removal and stemming with the Porter stemmer (Porter 1980).

Entity word: I also use the lemmas of the words that are part of e_1 and e_2.

Query word: Finally, I use the individual words that are part of the query string. This last feature is used for category C runs only (see below).

Participants in the competition were asked to classify their systems into categories A, B, C and D, depending on whether they use the manually annotated *WordNet* sense keys and/or the *Google* query:

A	WordNet = NO	Query = NO;
B	WordNet = YES	Query = NO;
C	WordNet = NO	Query = YES;
D	WordNet = YES	Query = YES.

I believe that having the target entities annotated with the correct *WordNet* sense keys is an unrealistic assumption for a real world application. Therefore, I use no *WordNet*-related features and I only experiment with conditions A and C.

As in section 4.9.4 above, I use a 1-nearest-neighbor classifier with a TF.IDF-weighted Dice coefficient. If the classifier makes no prediction (due to ties), I predict the majority class on the training data. Regardless of classifier's prediction, if the head nouns of the two entities e_1 and e_2 have the same lemma, I classify the example as negative.

In addition to accuracy, I use precision, recall and F-measure, defined as follows:

$$P = \Pr(\text{manual label} = \textsc{true} \mid \text{system guess} = \textsc{true})$$

$$R = \Pr(\text{system guess} = \textsc{true} \mid \text{manual label} = \textsc{true})$$

$$F = (2 \times P \times R)/(P + R)$$

$$Acc = \Pr(\text{system guess} = \text{manual label})$$

Tables 4.21 and 4.22 show the results for my type A and C experiments for different amounts of training data: 45 ($A1$, $C1$), 90 ($A2$, $C2$), 105 ($A3$, $C3$), and 140 ($A4$, $C4$) examples. All results are above the baseline: always propose the majority label in the test set: 'true' or 'false'. My category C results are slightly but consistently better than my category A results for all four evaluation scores (P, R, F, Acc), which suggests that knowing the query is helpful. Interestingly, systems $A2$ and $C2$ perform worse than $A1$ and $C1$, i.e., having more training data does not necessarily help with 1-nearest-neighbor classifiers. In fact, my category C system is the best-performing among the participating systems in *SemEval* task #4, and I have the third best results for category A.

Tables 4.23 and 4.24 show additional analysis for models $A4$ and $C4$. I study the effect of adding extra *Google* contexts (using up to 10 stars, rather than 8), and using different subsets of features. I show the results for the following experimental conditions: (a) 8 stars, leave-one-out cross-validation on the training data; (b) 8 stars, testing on the test data; and (c) 10 stars, testing on the test data. As the tables show, using 10 stars yields a slight overall improvement in accuracy for both $A4$ (from 65.4% to 67.3%) and $C4$ (from 67.0% to 68.1%). Both results represent improvements over the best $A4$ and $C4$ systems participating in *SemEval*, which achieved 66% and 67% accuracy, respectively.

Type	Train	Test	Relation	P	R	F	Accuracy
$A1$	35	80	Cause-Effect	58.2	78.0	66.7	60.00±10.95
	35	78	Instrument-Agency	62.5	78.9	69.8	66.67±11.03
	35	93	Product-Producer	77.3	54.8	64.2	59.14±10.16
	35	81	Origin-Entity	67.9	52.8	59.4	67.90±10.78
	35	71	Theme-Tool	50.0	31.0	38.3	59.15±11.62
	35	72	Part-Whole	51.9	53.8	52.8	65.28±11.52
	35	74	Content-Container	62.2	60.5	61.3	60.81±11.39
			Macro average	**61.4**	**58.6**	**58.9**	**62.7**
$A2$	70	80	Cause-Effect	58.0	70.7	63.7	58.75±10.95
	70	78	Instrument-Agency	65.9	71.1	68.4	67.95±10.99
	70	93	Product-Producer	80.0	77.4	78.7	72.04±9.86
	70	81	Origin-Entity	60.6	55.6	58.0	64.20±10.86
	70	71	Theme-Tool	45.0	31.0	36.7	56.34±11.57
	70	72	Part-Whole	41.7	38.5	40.0	58.33±11.53
	70	74	Content-Container	56.4	57.9	57.1	55.41±11.31
			Macro average	**58.2**	**57.5**	**57.5**	**61.9**
$A3$	105	80	Cause-Effect	62.5	73.2	67.4	63.75±10.94
	105	78	Instrument-Agency	65.9	76.3	70.7	69.23±10.94
	105	93	Product-Producer	75.0	67.7	71.2	63.44±10.14
	105	81	Origin-Entity	48.4	41.7	44.8	54.32±10.80
	105	71	Theme-Tool	62.5	51.7	56.6	67.61±11.54
	105	72	Part-Whole	50.0	46.2	48.0	63.89±11.54
	105	74	Content-Container	64.9	63.2	64.0	63.51±11.38
			Macro average	**61.3**	**60.0**	**60.4**	**63.7**
$A4$	140	80	Cause-Effect	63.5	80.5	71.0	66.25±10.89
	140	78	Instrument-Agency	70.0	73.7	71.8	71.79±10.83
	140	93	Product-Producer	76.3	72.6	74.4	66.67±10.07
	140	81	Origin-Entity	50.0	47.2	48.6	55.56±10.83
	140	71	Theme-Tool	61.5	55.2	58.2	67.61±11.54
	140	72	Part-Whole	52.2	46.2	49.0	65.28±11.52
	140	74	Content-Container	65.8	65.8	65.8	64.86±11.36
			Macro average	**62.7**	**63.0**	**62.7**	**65.4**
Baseline (majority)				81.3	42.9	30.8	57.0

Table 4.21: **Relations between nominals: evaluation results for systems of type A** Shown are the number of training and testing examples used, and the resulting precision, recall, F-measure and accuracy for each relation and micro-averaged over the 7 relations.

Type	**Train**	**Test**	**Relation**	**P**	**R**	**F**	**Accuracy**
$C1$	35	80	Cause-Effect	58.5	75.6	66.0	60.00±10.95
	35	78	Instrument-Agency	65.2	78.9	71.4	69.23±10.94
	35	93	Product-Producer	81.4	56.5	66.7	62.37±10.15
	35	81	Origin-Entity	67.9	52.8	59.4	67.90±10.78
	35	71	Theme-Tool	50.0	31.0	38.3	59.15±11.62
	35	72	Part-Whole	51.9	53.8	52.8	65.28±11.52
	35	74	Content-Container	62.2	60.5	61.3	60.81±11.39
			Macro average	**62.4**	**58.5**	**59.4**	**63.5**
$C2$	70	80	Cause-Effect	58.0	70.7	63.7	58.75±10.95
	70	78	Instrument-Agency	67.5	71.1	69.2	69.23±10.94
	70	93	Product-Producer	80.3	79.0	79.7	73.12±9.79
	70	81	Origin-Entity	60.6	55.6	58.0	64.20±10.86
	70	71	Theme-Tool	50.0	37.9	43.1	59.15±11.62
	70	72	Part-Whole	43.5	38.5	40.8	59.72±11.54
	70	74	Content-Container	56.4	57.9	57.1	55.41±11.31
			Macro average	**59.5**	**58.7**	**58.8**	**62.8**
$C3$	105	80	Cause-Effect	62.5	73.2	67.4	63.75±10.94
	105	78	Instrument-Agency	68.2	78.9	73.2	71.79±10.83
	105	93	Product-Producer	74.1	69.4	71.7	63.44±10.14
	105	81	Origin-Entity	56.8	58.3	57.5	61.73±10.89
	105	71	Theme-Tool	62.5	51.7	56.6	67.61±11.54
	105	72	Part-Whole	50.0	42.3	45.8	63.89±11.54
	105	74	Content-Container	64.9	63.2	64.0	63.51±11.38
			Macro average	**62.7**	**62.4**	**62.3**	**65.1**
$C4$	140	80	Cause-Effect	63.5	80.5	71.0	66.25±10.89
	140	78	Instrument-Agency	70.7	76.3	73.4	73.08±10.75
	140	93	Product-Producer	76.7	74.2	75.4	67.74±10.04
	140	81	Origin-Entity	59.0	63.9	61.3	64.20±10.86
	140	71	Theme-Tool	63.0	58.6	60.7	69.01±11.50
	140	72	Part-Whole	52.2	46.2	49.0	65.28±11.52
	140	74	Content-Container	64.1	65.8	64.9	63.51±11.38
			Macro average	**64.2**	**66.5**	**65.1**	**67.0**
Baseline (majority)				81.3	42.9	30.8	57.0

Table 4.22: **Relations between nominals: evaluation results for systems of type C** Shown are the number of training and testing examples used, and the resulting precision, recall, F-measure and accuracy for each relation and micro-averaged over the 7 relations.

Features Used	Leave-1-out: '8 *'	Test: '10 *'	Test: '8 *'
Cause-Effect			
sent	45.7	50.0±10.70	
p	55.0	53.75±10.85	
v	59.3	68.75±10.82	
$v+p$	57.1	63.75±10.94	
$v+p+c$	70.5	67.50±10.86	
$v+p+c+sent$ ($A4$)	58.5	66.25±10.89	66.25±10.89
$v+p+c+sent+query$ ($C4$)	59.3	66.25±10.89	66.25±10.89
Instrument-Agency			
sent	63.6	58.97±11.09	
p	62.1	70.51±10.89	
v	71.4	69.23±10.94	
$v+p$	70.7	70.51±10.89	
$v+p+c$	70.0	70.51±10.89	
$v+p+c+sent$ ($A4$)	68.6	71.79±10.83	71.79±10.83
$v+p+c+sent+query$ ($C4$)	70.0	73.08±10.75	73.08±10.75
Product-Producer			
sent	47.9	59.14±10.16	
p	55.7	58.06±10.15	
v	70.0	61.29±10.16	
$v+p$	66.4	65.59±10.10	
$v+p+c$	67.1	65.59±10.10	
$v+p+c+sent$ ($A4$)	66.4	69.89±9.96	66.67±10.07
$v+p+c+sent+query$ ($C4$)	67.9	69.89±9.96	66.67±10.07
Origin-Entity			
sent	64.3	72.84±10.55	
p	63.6	56.79±10.85	
v	69.3	71.60±10.62	
$v+p$	67.9	69.14±10.73	
$v+p+c$	66.4	70.37±10.68	
$v+p+c+sent$ ($A4$)	68.6	72.84±10.55	55.56±10.83
$v+p+c+sent+query$ ($C4$)	67.9	72.84±10.55	64.20±10.86

Table 4.23: **Relations between nominals: accuracy for different features and amount of Web data.** Shown is the accuracy for the following experimental conditions: (a) 8 stars, leave-one-out cross-validation on the training data; (b) 8 stars, testing on the test data; and (c) 10 stars, testing on the test data. (part 1)

Features Used	Leave-1-out: '8 *'	Test: '10 *'	Test: '8 *'
Theme-Tool			
sent	66.4	69.01±11.50	
p	56.4	56.34±11.57	
v	61.4	70.42±11.44	
$v + p$	56.4	67.61±11.54	
$v + p + c$	57.1	69.01±11.50	
$v + p + c + sent$ $(A4)$	52.1	61.97±11.63	67.61±11.54
$v + p + c + sent + query$ $(C4)$	52.9	61.97±11.63	69.01±11.50
sent	47.1	51.39±11.32	
p	57.1	54.17±11.43	
v	60.0	66.67±11.49	
$v + p$	62.1	63.89±11.54	
$v + p + c$	61.4	63.89±11.54	
$v + p + c + sent$ $(A4)$	60.0	61.11±11.55	65.28±11.52
$v + p + c + sent + query$ $(C4)$	60.0	61.11±11.55	65.28±11.52
Content-Container			
sent	56.4	54.05±11.27	
p	57.9	59.46±11.38	
v	71.4	67.57±11.30	
$v + p$	72.1	67.57±11.30	
$v + p + c$	72.9	67.57±11.30	
$v + p + c + sent$ $(A4)$	69.3	67.57±11.30	64.86±11.36
$v + p + c + sent + query$ $(C4)$	71.4	71.62±11.14	63.51±11.38
My average A4		**67.3**	**65.4**
Best avg. A4 on *SemEval*		**66.0**	
My average C4		**68.1**	**67.0**
Best avg. C4 on *SemEval*		**67.0**	

Table 4.24: **Relations between nominals: accuracy for different features and amount of Web data.** Shown is the accuracy for the following experimental conditions: (a) 8 stars, leave-one-out cross-validation on the training data; (b) 8 stars, testing on the test data; and (c) 10 stars, testing on the test data. (part 2)

4.10 Discussion

The verbal paraphrases I generate can be useful for a number of NLP tasks, e.g., for noun compound translation (in isolation) (Baldwin & Tanaka 2004), for paraphrase-augmented machine translation (Callison-Burch *et al.* 2006), for machine translation evaluation (Russo-Lassner *et al.* 2005; Kauchak & Barzilay 2006), for summarization evaluation (Zhou *et al.* 2006), etc. I will describe in chapter 5 an approach that integrates noun compound paraphrasing using prepositions as part of the machine translation process.

As I have shown above, assuming annotated training data, the paraphrasing verbs can be used as features in the prediction of abstract relations like `TIME` and `LOCATION`, which can be helpful for other applications. For example, Tatu & Moldovan (2005) achieve state-of-the-art results on the PASCAL Recognizing Textual Entailment challenge by making use of such abstract relations. See also appendix B for a detailed description of the task and for a discussion of how verbal paraphrases can be utilized directly.

In information retrieval, the paraphrasing verbs can be used for index normalization (Zhai 1997), query expansion, query refinement, results ranking, etc. For example, when querying for *migraine treatment*, pages containing good paraphrasing verbs like *relieve* or *prevent* could be preferred.

In data mining, the paraphrasing verbs can be used to seed a Web search that looks for particular classes of NPs (Etzioni *et al.* 2005), such as diseases, drugs, etc. For example, after having found that *prevent* is a good paraphrase for *migraine treatment*, I can use the query `"* which prevents migraines"` to obtain different treatments/drugs for migraine, e.g. *feverfew*, *Topamax*, *natural treatment*, *magnesium*, *Botox*, *Glucosamine*, etc. Using a different paraphrasing verb, e.g., using `"* reduces migraine"` can produce additional results: *lamotrigine*, *PFO closure*, *Butterbur Root*, *Clopidogrel*, *topamax*, *anti-convulsant*, *valproate*, *closure of patent foramen ovale*, *Fibromyalgia topamax*, *plant root extract*, *Petadolex*, *Antiepileptic Drug Keppra (Levetiracetam)*, *feverfew*, *Propranolol*, etc.

This is similar to the idea of a relational Web search of Cafarella *et al.* (2006), whose system TextRunner serves four types of relational queries, among which is one asking for all entities that are in a particular relation with a given entity, e.g., "*find all X such that X prevents migraines*".

4.11 Conclusion and Future Work

I have presented a simple unsupervised approach to noun-noun compound interpretation in terms of the predicates that can be used to paraphrase the hidden relation between the two nouns, which could be potentially useful for many NLP tasks. An important advantage of the approach is that it does not require knowledge about the meaning of the constituent nouns in order to correctly assign relations. A potential drawback is that it might not work well for low-frequency words.

I have further extended the method to measuring the semantic similarity between pairs of words, i.e. relational similarity. The evaluation of the approach on several relational similarity problems, including SAT verbal analogy, head-modifier relations, and relations between complex nominals has shown state-of-the-art performance.

The presented approach can be further extended to other combinations of parts of speech; not just noun-noun and adjective-noun. Using a full parser, or a dependency parser with a richer set of dependency features, e.g., as proposed by Padó & Lapata (2007), is another promising direction for future work.

Chapter 5

Improving Machine Translation with Noun Compound Paraphrases

This chapter describes an application of the methods developed in chapters 3 and 4 for noun compound paraphrasing to an important real-world task: *machine translation*. I propose a novel monolingual paraphrasing method based on syntactic transformations at the NP-level, which augments the training data with nearly equivalent sentence-level syntactic paraphrases of the original corpus, focused on the noun compounds. The idea is to recursively generate sentence variants where noun compounds are paraphrased using suitable prepositions, and vice-versa – NPs with an internal PP-attachment are turned into noun compounds. The evaluation results show an improvement equivalent to 33%-50% of that of doubling the amount of training data. The general idea of the approach was described in (Nakov & Hearst 2007).

5.1 Introduction

Most modern Statistical Machine Translation systems rely on aligned sentences of bilingual corpora for training, from which they learn how to translate small pieces of text. In many cases, the learned pieces are equivalent but syntactically different from the text given at translation time, and the potential for a high-quality translation is missed.

In this section, I describe a method for expanding the training set with conceptually similar but syntactically differing paraphrases. I paraphrase source language side training sentences, focusing on noun phrases (NPs) and noun compounds (NCs), which have been previously shown to be very frequent: Koehn & Knight (2003) observe that roughly half of the words in news texts are covered by NPs/PPs, and Baldwin & Tanaka (2004) report that 2.6% of the tokens in the *British National Corpus* and 3.9% of the tokens in the *Reuters corpus* are covered by noun compounds.

The proposed approach is novel in that it augments the training corpus with paraphrases of the original sentences, thus increasing the training set without increasing the number of training translation pairs needed. It is also monolingual; other related approaches map from the source language to other languages in order to obtain paraphrases. And it is general enough to be domain independent, although the paraphrasing rules I use are currently English-specific.

In the experiments below, I show that the proposed monolingual paraphrasing approach can improve *Bleu* scores (Papineni *et al.* 2001) on small training datasets when translating from English to Spanish.

5.2 Related Work

Recent work in automatic corpus-based paraphrasing includes using bilingual corpora to find alternative expressions for the same term, e.g., (Barzilay & McKeown 2001) and (Pang *et al.* 2003), or multiple expressions of the same concept in one language, as in newswire text (Shinyama *et al.* 2002).

Recently, paraphrases have been used to improve the *evaluation* of machine translation systems. Kauchak & Barzilay (2006) argue that automated evaluation measures like *Bleu* end up comparing n-gram overlap rather than semantic similarity to a reference text. They performed an experiment asking two human translators to translate 10,000 sentences, and found that less than 0.2% of the translations were identical, and 60% differed by more than ten words. Therefore, they proposed an evaluation method that includes paraphrasing of the machine-produced translations, and demonstrated improved correlation with human judgements compared to *Bleu*. In a similar spirit, Zhou *et al.* (2006) use a paraphrase table extracted from a bilingual corpus to improve evaluation of automatic summarization algorithms.

My approach is most closely related to that of Callison-Burch *et al.* (2006), who translate English sentences into Spanish and French by trying to substitute unknown source phrases with suitable paraphrases. The paraphrases are extracted using the bilingual method of Bannard & Callison-Burch (2005), which finds paraphrases in one language using a phrase in a second language as a pivot. For example, if in a parallel English-German corpus, the English phrases *under control* and *in check* happen to be aligned (in different sentences) to the same German phrase *unter controlle*, they would be hypothesized to be paraphrases of each other with some probability. Callison-Burch *et al.* (2006) extract English paraphrases using as pivots eight additional languages from the *Europarl* corpus (Koehn 2005). These paraphrases are then incorporated in the machine translation process by adding them as additional entries in the English-Spanish phrase table and pairing them with the foreign translation of the original phrase. Finally, the system is tuned using minimum error rate training (Och 2003) with an extra feature penalizing the low probability paraphrases. The system achieved dramatic increases in coverage (from 48% to 90% of the test word types when using 10,000 training sentence pairs), and notable increase on *Bleu* (up to 1.5 points). However, the method requires large multi-lingual parallel corpora, which makes it domain-dependent and most likely limits its source language to Chinese, Arabic, and the languages of the EU.

Another important related research effort is in translating units of text smaller than a sentence, e.g., noun-noun compounds (Grefenstette 1999; Tanaka & Baldwin 2003; Baldwin & Tanaka 2004), noun phrases (Cao & Li 2002; Koehn & Knight 2003), named entities (Al-Onaizan & Knight 2001), and technical terms (Nagata *et al.* 2001). Although I focus on paraphrasing NPs, unlike the work above, I paraphrase and translate full sentences, as opposed to translating NPs as a stand-alone component.

5.3 Method

I propose a novel approach for improving statistical machine translation using monolingual paraphrases. Given a sentence from the source (English) side of the training corpus, I generate conservative meaning-preserving syntactic paraphrases of the sentence,

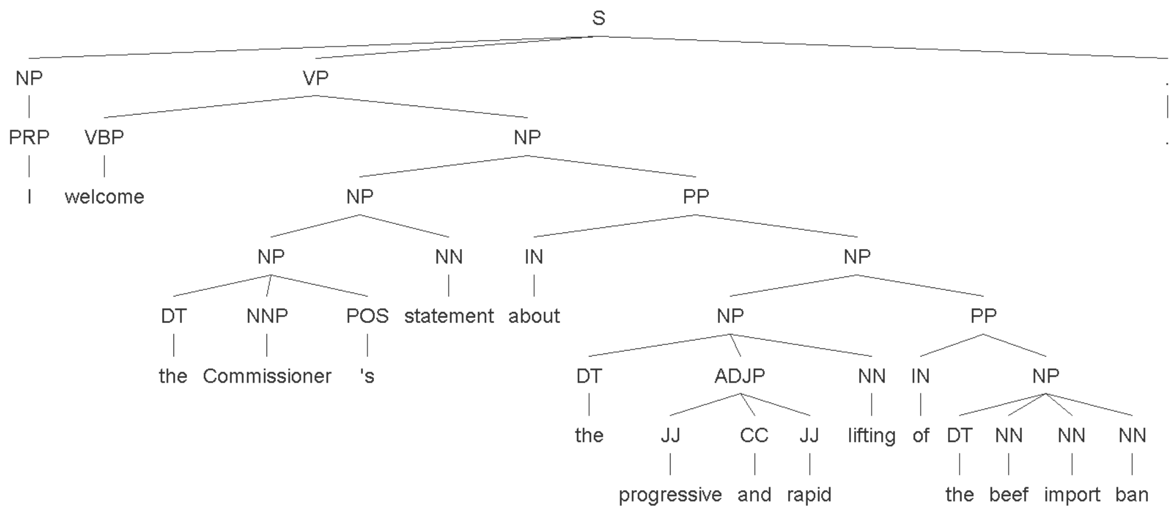

Figure 5.1: **Sample parse tree generated by the *Stanford parser*.** I transform noun compounds into NPs with an internal PP-attachment; I turn NPs with an internal PP-attachment into noun compounds, or into NPs with an internal possessive marker; and I remove possessive markers whenever possible, or substitute them with *of*. All these transformations are applied recursively.

and I append these paraphrased sentences to the training corpus. Each paraphrased sentence is paired with the foreign (Spanish) translation that is associated with the original sentence in the training data. This augmented training corpus can then be used to train a statistical machine translation (SMT) system.

I also introduce a variation on this idea that can be used with a phrase-based SMT. In this alternative, the source-language phrases from the phrase table are paraphrased, but again using the source language phrases for the paraphrasing, as opposed to using foreign aligned corpora to modify the phrases, as done by Callison-Burch *et al.* (2006). I also combine these approaches.

I welcome the Commissioner 's statement about the progressive and rapid beef import ban lifting .
I welcome the progressive and rapid beef import ban lifting Commissioner 's statement .
I welcome the Commissioner 's statement about the beef import ban 's progressive and rapid lifting .
I welcome the beef import ban 's progressive and rapid lifting Commissioner 's statement .
I welcome the Commissioner 's statement about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the Commissioner statement about the progressive and rapid lifting of the beef import ban .
I welcome the Commissioner statement about the progressive and rapid beef import ban lifting .
I welcome the progressive and rapid beef import ban lifting Commissioner statement .
I welcome the Commissioner statement about the beef import ban 's progressive and rapid lifting .
I welcome the beef import ban 's progressive and rapid lifting Commissioner statement .
I welcome the Commissioner statement about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the statement of Commissioner about the progressive and rapid lifting of the beef import ban .
I welcome the statement of Commissioner about the progressive and rapid beef import ban lifting .
I welcome the statement of Commissioner about the beef import ban 's progressive and rapid lifting .
I welcome the statement of Commissioner about the progressive and rapid lifting of the *ban on beef imports* .
I welcome the statement of the Commissioner about the progressive and rapid lifting of the beef import ban .
I welcome the statement of the Commissioner about the progressive and rapid beef import ban lifting .
I welcome the statement of the Commissioner about the beef import ban 's progressive and rapid lifting .
I welcome the statement of the Commissioner about the progressive and rapid lifting of the *ban on beef imports* .
The EU budget , as an instrument of economic policy , amounts to 1.25 % of European GDP .
The EU budget , as an economic policy instrument , amounts to 1.25 % of European GDP .
The EU budget , as an economic policy 's instrument , amounts to 1.25 % of European GDP .
The *EU 's budget* , as an instrument of economic policy , amounts to 1.25 % of European GDP .
The *EU 's budget* , as an economic policy instrument , amounts to 1.25 % of European GDP .
The *EU 's budget* , as an economic policy 's instrument , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an instrument of economic policy , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an economic policy instrument , amounts to 1.25 % of European GDP .
The *budget of the EU* , as an economic policy 's instrument , amounts to 1.25 % of European GDP .
We must cooperate internationally , and this should include UN initiatives .
We must cooperate internationally , and this should include *initiatives of the UN* .
We must cooperate internationally , and this should include *initiatives at the UN* .
We must cooperate internationally , and this should include *initiatives in the UN* .
Both reports on economic policy confirm the impression that environment policy is only a stepchild .
Both reports on economic policy confirm the impression that *policy on the environment* is only a stepchild .
Both reports on economic policy confirm the impression that *policy on environment* is only a stepchild .
Both reports on economic policy confirm the impression that *policy for the environment* is only a stepchild .
Both economic policy reports confirm the impression that environment policy is only a stepchild .
Both economic policy reports confirm the impression that *policy on the environment* is only a stepchild .
Both economic policy reports confirm the impression that *policy on environment* is only a stepchild .
Both economic policy reports confirm the impression that *policy for the environment* is only a stepchild .

Table 5.1: **Sample sentences and their automatically generated paraphrases.** Paraphrased noun compounds are in italics.

5.4 Paraphrasing

Consider the sentence:

> "*I welcome the Commissioner's statement about the progressive and rapid lifting of the beef import ban.*".

The corresponding syntactic parse tree, produced using the *Stanford syntactic parser* (Klein & Manning 2003), is shown in Figure 5.1. From this parse, I generate paraphrases using the six syntactic transformations shown below, which are performed recursively. Table 5.1 shows some sentences and the corresponding paraphrases for them, including all the paraphrases I generate from the parse tree in Figure 5.1.

1. $[_{\mathbf{NP}}\ \mathbf{NP}_1\ \mathbf{P}\ \mathbf{NP}_2] \Rightarrow [_{\mathbf{NP}}\ \mathbf{NP}_2\ \mathbf{NP}_1]$**.**

   the lifting of the beef import ban ⇒ *the beef import ban lifting*

2. $[_{\mathbf{NP}}\ \mathbf{NP}_1\ \boldsymbol{of}\ \mathbf{NP}_2] \Rightarrow [_{\mathbf{NP}}\ \mathbf{NP}_2\ \mathbf{poss}\ \mathbf{NP}_1]$**.**

   the lifting of the beef import ban ⇒ *the beef import ban's lifting*

3. $\mathbf{NP}_{poss} \Rightarrow \mathbf{NP}$**.**

   Commissioner's statement ⇒ *Commissioner statement*

4. $\mathbf{NP}_{poss} \Rightarrow \mathbf{NP}_{PP_{of}}$**.**

   Commissioner's statement ⇒ *statement of (the) Commissioner*

5. $\mathbf{NP}_{NC} \Rightarrow \mathbf{NP}_{poss}$**.**

   inquiry committee chairman ⇒ *inquiry committee's chairman*

6. $\mathbf{NP}_{NC} \Rightarrow \mathbf{NP}_{PP}$**.**

   the beef import ban ⇒ *the ban on beef import* where:

poss	possessive marker: ' or 's;
P	preposition;
$\mathbf{NP}_{PP}$	NP with an internal PP-attachment;
$\mathbf{NP}_{PP_{of}}$	NP with an internal PP headed by *of*;
$\mathbf{NP}_{poss}$	NP with an internal possessive marker;
$\mathbf{NP}_{NC}$	NP that is a noun compound.

In order to prevent transformations (1) and (2) from constructing awkward NPs, I impose certain limitations on NP_1 and NP_2. They cannot span a verb, a preposition or a quotation mark (although they can contain some kinds of nested phrases, e.g., an ADJP in case of coordinated adjectives, as in *the progressive and controlled lifting*). Thus, the phrase *reduction in the taxation of labour* is not transformed into *taxation of labour reduction* or *taxation of labour's reduction*. I further require the head to be a noun and do not allow it to be an indefinite pronoun like *anyone*, *everybody*, and *someone*.

Transformations (1) and (2) are more complex than they may look. In order to be able to handle some hard cases, I apply additional restrictions and transformations.

First, some determiners, pre-determiners and possessive adjectives must be eliminated in case of conflict between NP_1 and NP_2, e.g., *the lifting of this ban* can be paraphrased as *the ban lifting*, but not as *this ban's lifting*.

Second, in case both NP_1 and NP_2 contain adjectives, these adjectives have to be put in the right order, e.g., *the first statement of the new commissioner* can be paraphrased as *the first new commissioner's statement*, but not *the new first commissioner's statement*. There is also the option of not re-ordering them, e.g., *the new commissioner's first statement*.

Third, further complications are due to scope ambiguities of modifiers of NP_1. For example, in *the first statement of the new commissioner*, the scope of the adjective *first* is not *statement* alone, but *statement of the new commissioner*. This is very different for the NP *the biggest problem of the whole idea*, where the adjective *biggest* applies to *problem* only, and therefore it cannot be transformed to *the biggest whole idea's problem* (although I do allow for *the whole idea's biggest problem*).

While the first four transformations are purely syntactic, (5) and (6) are not. The algorithm must determine whether a possessive marker is feasible for (5) and must choose the correct preposition for (6). In both cases, for noun compounds of length three or more, I also need to choose the correct position to modify, e.g., *inquiry's committee chairman* vs. *inquiry committee's chairman*.

In order to ensure accuracy of the paraphrases, I use a variation of the Web-based approaches presented in chapter 3, generating and testing the paraphrases in the context of the preceding and the following word in the sentence. First, I split the noun compound into two sub-parts N_1 and N_2 in all possible ways, e.g., *beef import ban lifting* would be split as: (a) N_1="*beef*", N_2="*import ban lifting*", (b) N_1="*beef import*", N_2="*ban lifting*", and (c) N_1="*beef import ban*", N_2="*lifting*". For each split, I issue exact phrase queries to *Google* using the following patterns:

"lt N_1 poss N_2 rt"

"lt N_2 prep det N_1' rt"

"lt N_2 that be det N_1' rt"

"lt N_2 that be prep det N_1' rt"

where:

N_1' can be a singular or a plural form of N_1;

`lt` is the word preceding N_1 in the original sentence or empty if none;

`rt` is the word following N_2 in the original sentence or empty if none;

`poss` is a possessive marker ('s or ');

`that` is *that*, *which* or *who*;

`be` is *is* or *are*;

`det` is a determiner (*the*, *a*, *an*, or none);

`prep` is one of the eight prepositions used by Lauer (1995) for noun compound interpretation: *about*, *at*, *for*, *from*, *in*, *of*, *on*, and *with*.

Given a split, I collect the number of page hits for each instantiation of the above paraphrase patterns filtering out the ones whose page hit counts are less than ten. I then calculate the total number of page hits H for all paraphrases (for all splits and all patterns), and I retain the ones whose page hits counts are at least 10% of H, which allows for multiple paraphrases (possibly corresponding to different splits) for a given noun compound. If no paraphrases are retained, I repeat the above procedure with `lt` set to the empty string. If there are still no good paraphrases, I set `rt` to the empty string. If this does not help either, I make a final attempt, by setting both `lt` and `rt` to the empty string.

The paraphrased NCs are shown in italics in Table 5.1, e.g., *EU budget* is paraphrased as *EU's budget* and *budget of the EU*; also *environment policy* becomes *policy on environment*, *policy on the environment*, and *policy for the environment*; and *UN initiatives* is paraphrased as *initiatives of the UN*, *initiatives at the UN*, and *initiatives in the UN*.

1	**% of members of the irish parliament** % of irish parliament members % of irish parliament 's members
2	**universal service of quality .** universal quality service . quality universal service . quality 's universal service .
3	**action at community level** community level action
4	**, and the aptitude for communication and** , and the communication aptitude and
5	**to the fall-out from chernobyl .** to the chernobyl fall-out .
6	**flexibility in development - and quick** development flexibility - and quick
7	**, however , the committee on transport** , however , the transport committee
8	**and the danger of infection with aids** and the danger of aids infection and the aids infection danger and the aids infection 's danger

Table 5.2: **Sample English phrases from the phrase table and corresponding automatically generated paraphrases.**

I apply the same algorithm to paraphrasing English *phrases* from the phrase table, but without transformations (5) and (6). See Table 5.2 for sample paraphrases.

5.5 Evaluation

I trained and evaluated several English to Spanish phrase-based SMT systems using the standard training, development and test sets of the *Europarl* corpus (Koehn 2005).

First, I generate two directed word-level alignments, English-Spanish and Spanish-English, using IBM model 4 (Brown *et al.* 1993), and I combine them using the *interect+grow heuristic* described in Och & Ney (2003). Then I extract phrase-level translation pairs using the *alignment template approach* (Och & Ney 2004). The set of pairs of English phrases and their translations into Spanish form a phrase table where each pair is associated with five parameters: a forward phrase translation probability, a reverse phrase translation probability, a forward lexical translation probability, a reverse lexical translation probability, and a phrase penalty.

I train a log-linear model using the standard set of feature functions of PHARAOH (Koehn 2004): language model probability, word penalty, distortion cost, and the five parameters from the phrase table. I set the feature weights by optimizing the *Bleu* score directly using *minimum error rate training* (Och 2003) on the first 500 sentences from the development set. Then, I use the PHARAOH beam search decoder (Koehn 2004) in order to produce the translations for the 2,000 test sentences. Finally, these translations are compared to the gold standard test set using *Bleu*.

Using the above procedure, I build and evaluate a baseline system S, trained on the original training corpus. The second system I build, S_{parW}, uses a version of the training corpus which is augmented with syntactic paraphrases of the sentences from the English side of the training corpus paired with the corresponding Spanish translations of the original English sentences. I also build a new system S_{par}, which excludes syntactic transformations (5) and (6) from the paraphrasing process in order to see the impact of not breaking noun compounds and not using the Web.

System $S^{\star}$ paraphrases and augments the phrase table of the baseline system S using syntactic transformations (1)-(4), similarly to S_{par}, i.e. without paraphrasing noun compounds. Similarly, $S^{\star}_{parW}$ is obtained by paraphrasing the phrase table of S_{parW}.

Finally, I merge the phrase tables for some of the above systems, which I designate with a "+", e.g., $S + S_{parW}$ and $S^\star + S^\star_{parW}$. In these merges, the phrases from the first phrase table are given priority over those from the second one, in case a phrase pair is present in both phrase tables. This is important since the parameters estimated from the original corpus are more reliable.

Following Bannard & Callison-Burch (2005), I also perform an experiment with an additional feature F_{parW} for each phrase: Its value is 1 if the phrase is in the phrase table of S, and 0.5 if it comes from the phrase table of S_{parW}. As before, I optimize the weights using minimum error rate training. For $S^\star + S^\star_{parW}$, I also try using two features: in addition to F_{parW}, we I introduce $F_\star$, whose value is 0.5 if the phrase comes from paraphrasing a phrase table entry, and 1 if was in the original phrase table.

5.6 Results and Discussion

The results of the evaluation are shown in Tables 5.3 and 5.5. I use the notation shown in Table 5.4.

The differences between the baseline and the remaining systems shown in Table 5.3 are statistically significant: tested using the bootstrapping method described in Zhang & Vogel (2004) with 2,000 samples.

Gain of 33%–50% compared to doubling the training data. As Table 5.5 shows, neither paraphrasing the training sentences, S_{parW}, nor paraphrasing the phrase table, $S^\star$, lead to notable improvements. For 10k training sentences the two systems are comparable and improve *Bleu* by 0.3, for 40k sentences, $S^\star$ performs as the baseline

System	Bleu	n-gram precision				Bleu		# of phrases	
		1-gr.	2-gr.	3-gr.	4-gr.	BP	ration	gener.	used
S (baseline)	22.38	55.4	27.9	16.6	10.0	0.995	0.995	180,996	40,598
S_{par}	21.89	55.7	27.8	16.5	10.0	0.973	0.973	192,613	42,169
S_{parW}	22.57	55.1	27.8	16.7	10.2	1.000	1.000	202,018	42,523
$S^{\star}$	22.58	55.4	28.0	16.7	10.1	1.000	1.001	206,877	40,697
$S + S_{par}$	22.73	55.8	28.3	16.9	10.3	0.994	0.994	262,013	54,486
$S + S_{parW}$	**23.05**	55.8	28.5	17.1	10.6	0.995	0.995	280,146	56,189
$S + S_{parW}\dagger$	**23.13**	55.8	28.5	17.1	10.5	1.000	1.002	280,146	56,189
$S^{\star} + S^{\star}_{parW}$	**23.09**	56.1	28.7	17.2	10.6	0.993	0.993	327,085	56,417
$S^{\star} + S^{\star}_{parW}\ddagger$	**23.09**	55.8	28.4	17.1	10.5	1.000	1.001	327,085	56,417

Table 5.3: ***Bleu*** **scores and n-gram precisions (in %s) for 10k training sentences**. The last two columns show the total number of phrase pairs in the phrase table and the number usable on the test set, respectively. See the text for a description of the systems.

S	baseline, trained on the original corpus;
S_{par}	original corpus, augmented with sentence-level paraphrases, no transformations (5) and (6) (i.e. without using the Web);
S_{parW}	original corpus, augmented with sentence-level paraphrases, all transformations;
$\star$	means paraphrasing the phrase table;
$+$	means merging the phrase tables;
$\dagger$	using an extra feature: F_{parW};
$\ddagger$	using two extra features: $F_{\star}$, F_{parW}.

Table 5.4: **Notation for the experimental runs.**

System	# of training sentences			
	10k	20k	40k	80k
S (baseline)	22.38	24.33	26.48	27.05
S_{parW}	22.57	24.41	25.96	
$S^{\star}$	22.58	25.00	26.48	
$S + S_{parW}$	**23.05**	**25.01**	**26.75**	

Table 5.5: ***Bleu*** **scores for different number of training sentences.**

and S_{parW} even drops below it. However, when merging the phrase tables of S and S_{parW}, I get an improvement of almost 0.7 for 10k and 20k sentences, and about 0.3 for 40k sentences. While this improvement might look small, it is comparable to that of Bannard & Callison-Burch (2005), who obtain 0.7 improvement for 10k sentences and 1.0 for 20k sentences (but translating in the reverse direction: from Spanish into English). Note also, that the 0.7 *Bleu* improvement for 10k and 20k sentences is about 1/3 of the 2 *Bleu* points obtained by the baseline system by doubling the training size. Note also that the 0.3 gain on *Bleu* for 40k sentences is equal to half of what will be gained if I train on 80k sentences.

Improved precision for all n-grams. Table 5.3 provides a comparison of different systems trained on 10k sentences. In addition to the *Bleu* score, I also give its elements: n-gram precisions, BP (brevity penalty) and ration. Comparing the baseline with the last four systems, we can see that unigram, bigram, trigram and fourgram precisions are all improved by between 0.4% and 0.7%.

Importance of noun compound splitting. System S_{par} is trained on the training corpus augmented with paraphrased sentences, where the noun compound splitting transformations (5) and (6) are not used, i.e. the paraphrases are purely syntactic and use no Web counts. We can see that omitting these rules causes the results to go below the baseline: while there is a 0.3% gain on unigram precision, bigram and trigram precision go down by about 0.1%. More importantly, BP goes down as well: since the sentence-level paraphrases (except for the possessives which are infrequent) mainly convert NPs into noun compounds, the resulting sentences are shorter, which causes the translation model to learn to generate shorter sentences. Note that this effect is not observed for S_{parW}, where transformations (5) and (6) make the sentences longer, and therefore balance the BP to be exactly 1.0. A somewhat different kind of argument applies to $S+S_{par}$, which is worse than $S+S_{parW}$, but not because of BP. In this case, there is no improvement for unigrams, but a consistent 0.2-0.3% drop for bigrams, trigrams and fourgrams. The reason for this is shown in the last column of Table 5.5 – omitting the noun compound splitting transformations (5) and (6) results in fewer training sentences, which means fewer phrases in the phrase table and consequently fewer phrases compatible with the test set.

More usable phrases. The last two columns of Table 5.3 show that more phrases in the phrase table mean an increased number of usable phrases as well. A notable exception is $S^\star$: its phrase table is bigger than those of S_{par} and S_{parW}, but it contains fewer phrases compatible with the test set then them (but still more than the baseline). This suggests that additional phrases extracted from paraphrased sentences are more likely to be usable at test time than additional phrases generated by paraphrasing the phrase table.

Paraphrasing the phrase table vs. paraphrasing the training corpus. As Tables 5.5 and 5.3 show, paraphrasing the phrase table $S^\star$ (*Bleu* score 22.58%) does not compete against paraphrasing the training corpus followed by merging the resulting phrase table with the phrase table for the original corpus[1], as in $S + S_{parW}$ (*Bleu* score 23.05%). I also try to paraphrase the phrase table of $S + S_{parW}$, but the resulting system $S^\star + S^\star_{parW}$ yields little improvement: 23.09% *Bleu* score. Adding the two extra features, $F_\star$ and F_{parW}, does not yield improvements as well: $S^\star + S^\star_{parW}\ddagger$ achieves the same *Bleu*

[1]Note that $S^\star$ does not use syntactic transformations (5) and (6). However, as the results for $S + S_{par}$ show, the claim still holds even if transformations (5) and (6) are also excluded when paraphrasing the sentences: the *Bleu* score for $S + S_{par}$ is 22.73% vs. 22.58% for $S^\star$.

score as $S^{\star} + S^{\star}_{parW}$. This shows that extracting additional phrases from the augmented corpus is a better idea than paraphrasing the phrase table, which can result in erroneous splitting of noun phrases. Paraphrasing whole sentences as opposed to paraphrasing the phrase table could potentially improve the approach of Callison-Burch *et al.* (2006): while low probability and context dependency could be problematic, a language model could help filtering the bad sentences out. Such filtering could potentially improve my results as well. Finally, note that different paraphrasing strategies could be used when paraphrasing phrases vs. sentences. For example, paraphrasing the phrase table can be done more aggressively: if an ungrammatical phrase is generated in the phrase table, it will probably have no negative effect on translation quality since it will be unlikely to occur in the test data.

Quality of the paraphrases. An important difference between my syntactic paraphrasing method and the multi-lingual approach of Callison-Burch *et al.* (2006) is that their paraphrases are only contextually synonymous and often depart significantly from the original meaning. As a result, they could not achieve improvements by simply augmenting the phrase table: this introduces too much noise and yields an accuracy significantly below the baseline – by 3-4% on *Bleu*. In order to achieve an improvement, they had to introduce an extra feature penalizing the low probability paraphrases and promoting the original phrases in the phrase table. In contrast, my paraphrases are meaning preserving and context-independent: introducing the feature F_{parW}, which penalizes phrases coming from the paraphrased corpus, in system $S + S_{parW}\dagger$ yields a tiny improvement on *Bleu* score (23.13% vs. 23.05%), i.e., the phrases extracted from my augmented corpus are almost as good as the ones from the original corpus. Finally, note that my paraphrasing method is complementary to that of Callison-Burch *et al.* (2006) and therefore the two can be combined: the strength of my approach is at improving the coverage of longer phrases using syntactic paraphrases, while the strength of theirs is at improving the vocabulary coverage with extra words extracted from additional corpora (although they do get some gain from using longer phrases as well).

Translating into English. I tried translating in the reverse direction, paraphrasing the target (English) language side, which resulted in decreased performance. This is not surprising: the set of available source phrases remains the same, and a possible improvement could potentially come from producing a more fluent translation only, e.g., from turning an NP with an internal PP into a noun compound. However, unlike the original translations, the extra ones are less likely to be judged correct since they were not observed in training.

Problems and limitations. Error analysis suggests that the major problems for the proposed paraphrasing method are caused by incorrect PP-attachments in the parse tree. Somewhat less frequent and therefore less important are errors of POS tagging. At present I use a parser, which limits the applicability of the method to languages for which syntactic parsers are available. In fact, the kinds of paraphrases I use are simple and can be approximated by using a shallow parser or just a POS tagger; this would lead to errors of PP-attachment, but these attachments are often assigned incorrectly by parsers anyway.

The central target of my paraphrases are noun compounds – I turn an NP with an internal PP into a noun compound and vice versa – which limits its applicability of the approach to languages where noun compounds are a frequent phenomenon, e.g., Germanic, but not Romance or Slavic languages.

From a practical viewpoint, a limitation of the method is that it increases the size of the phrase table and/or of the training corpus, which slows down the processes of both training and translation, and limits its applicability to relatively small corpora for computational reasons.

Finally, as Table 5.5 shows, the improvements get smaller for bigger training corpora, which suggests that it gets harder to generate useful paraphrases that are not already present in the corpus.

5.7 Conclusion and Future Work

I have presented a novel domain-independent approach for improving statistical machine translation by augmenting the training corpus with monolingual paraphrases of the source language side sentences, thus increasing the training data "for free", by creating it from data that is already available rather than having to create more aligned data.

While in my experiments I use phrase-based SMT, any other MT approach that learns from parallel corpora could potentially benefit from the proposed syntactic corpus augmentation idea. At present, my paraphrasing rules are English-specific, but they could be easily adapted to other Germanic languages, which make heavy use of noun compounds; the general idea of automatically generating nearly equivalent source side syntactic paraphrases can in principle be applied to any language. The current version of the method can be considered preliminary, as it is limited to NPs; still, the results are already encouraging, and the approach is worth considering when building MT systems from small corpora, e.g., in case of resource-poor language pairs, in specific domains, etc.

Better use of the Web could be made for paraphrasing noun compounds (e.g., using verbal paraphrases as in chapters 3 and 4), other syntactic transformations could be tried (e.g., adding/removing complementizers like *that*, which are not obligatory in English, or adding/removing commas from non-obligatory positions), and a language model could be used to filter out the bad paraphrases.

Even more promising, but not that simple, seems using a tree-to-tree syntax-based SMT system and learning transformations that can make the source language trees structurally closer to the target-language ones, e.g., the English sentence "*Remember the guy who you are <u>with</u>!*" would be transformed to "*Remember the guy <u>with</u> whom you are!*", whose word order is closer to Spanish "*¡Recuerda al individuo <u>con</u> quien estás!*", which might facilitate the translation process.

Finally, the process could be made part of the decoding, which would eliminate the need of paraphrasing the training corpus and might allow the dynamic generation of paraphrases both for the phrase table and for the test sentence.

Chapter 6

Other Applications

This chapter describes applications of the methods developed in chapter 3: to two important structural ambiguity problems a syntactic parser faces – *prepositional phrase attachment* and *NP coordination*. Using word-association scores, Web-derived surface features, and paraphrases, I achieve 84% accuracy for the former task and 80% for the latter one, which is on par with the state-of-the-art. A shorter version of this work appeared in (Nakov & Hearst 2005c).

6.1 Prepositional Phrase Attachment

6.1.1 Introduction

A long-standing challenge for syntactic parsers is the attachment decision for prepositional phrases. In a configuration where a verb takes a noun complement that is followed by a prepositional phrase (PP), the problem arises of whether the prepositional phrase attaches to the noun or to the verb. Consider for example the following contrastive pair of sentences:

(1)	*Peter spent millions of dollars.*	(noun)
(2)	*Peter spent time with his family.*	(verb)

In the first sentence, the prepositional phrase *millions of dollars* attaches to the noun *millions*, while in the second sentence the prepositional phrase *with his family* attaches to the verb *spent*.

Below I will address a formulation of the PP-attachment problem that casts these associations out of context, as the quadruple (v, n_1, p, n_2), where v is the verb, n_1 is the head of the direct object, p is the preposition (the head of the PP), and n_2 is the head of the noun phrase inside the prepositional phrase. For example, the quadruple for (2) is (*spent*, *time*, *with*, *family*).

6.1.2 Related Work

While most early work on PP-attachment ambiguity resolution relied primarily on deterministic pragmatic and syntactic considerations (e.g., *minimal attachment*, *right association*, etc.), recent research on the problem is dominated by probabilistic machine learning approaches: supervised and unsupervised.

Supervised Approaches

Ratnaparkhi *et al.* (1994b) created a dataset of 27,937 (20,801 training; 4,039 development; 3,097 testing) quadruples (v, n_1, p, n_2) from the *Wall Street Journal*; I will refer to it as *Ratnaparkhi's dataset*. They achieved 81.6% accuracy with a maximum entropy classifier and a binary hierarchy of word classes derived using mutual information. They also found the human performance for the task to be 88% without context, and 93% when sentence context is provided.

Ratnaparkhi's dataset has been used as the benchmark dataset for this task. Collins & Brooks (1995) propose a supervised back-off model, which achieves 84.5% accuracy. Zavrel *et al.* (1997) use memory-based learning and various similarity metrics yielding 84.4% accuracy. Stetina & Nagao (1997) use a supervised method with decision trees and *WordNet* classes which achieves 88.1% accuracy. Abney *et al.* (1999) report 84.6% accuracy using boosting. Using SVMs with weighted polynomial information gain kernels, Vanschoenwinkel & Manderick (2003) achieve 84.8% accuracy. Alegre *et al.* (1999) report 86% with neural networks. Using a nearest-neighbor classifier, Zhao & Lin (2004) achieve 86.5% accuracy. Toutanova *et al.* (2004) report 87.5% accuracy using *WordNet* and morphological and syntactic analysis.

On their own dataset of 500 test examples, Brill & Resnik (1994) use supervised transformation-based learning and conceptual classes derived from *WordNet*, achieving 82% accuracy. On a different dataset, Li (2002) achieves 88.2% accuracy with word clustering based on co-occurrence data. Olteanu & Moldovan (2005) use SVM with a large number of complex syntactic and semantic features, and n-gram frequencies from the Web. They could not use *Ratnaparkhi's dataset*, since their formulation of the problem is a bit different: they use features extracted from the whole sentence, as opposed to using the quadruple only. They achieve 93.62% on a dataset extracted from the *Penn Treebank*; on another dataset derived from *FrameNet* (Baker *et al.* 1998), they achieve 91.79% and 92.85%, depending on whether manually annotated semantic information is made available to the model.

Unsupervised Approaches

Unsupervised approaches make attachment decisions using co-occurrence statistics drawn from text collections. The idea was pioneered by Hindle & Rooth (1993), who use a partially parsed corpus to calculate lexical associations over subsets of the tuple (v, n_1, p) (e.g., comparing $\Pr(p|n_1)$ to $\Pr(p|v)$), achieving 80% accuracy at 80% coverage.

Ratnaparkhi (1998) develops an unsupervised method that collects statistics from text annotated with POS tags and morphological base forms. An extraction heuristic is used to identify unambiguous attachment decisions, for example, the algorithm can assume a *noun attachment* if there is no verb within k words to the left of the preposition in a given sentence, among other conditions. In his experiments, this heuristic uncovers 910K unique tuples of the form (v, p, n_2) and (n, p, n_2), which suggest the correct attachment only about 69% of the time. The tuples are used by classifiers, the best of which achieved 81.9% accuracy on *Ratnaparkhi's dataset*. Note that these classifiers are not trained on real training data, but on the output of an unsupervised method, i.e. the method stays unsupervised.

Pantel & Lin (2000) describe an unsupervised method, which uses a collocation database, a thesaurus, a dependency parser, and a large corpus (125M words), achieving 84.31% accuracy on the same dataset.

Recently, there has been a growing interest in using n-gram frequencies derived from the Web, as opposed to estimating them from a static corpus. Volk (2000) used Web-derived n-gram frequencies to solve the PP-attachment problem for German. He compared $\Pr(p|n_1)$ to $\Pr(p|v)$, estimated as $\Pr(p|x) = \#(x, p)/\#(x)$, where x can be n_1 or v, and the frequencies are obtained from *Altavista* using the `NEAR` operator. The approach was able to make a decision for 58% of his examples with a accuracy of 75% (baseline 63%). Volk (2001) later improved on these results by comparing $\Pr(p, n_2|n_1)$ to $\Pr(p, n_2|v)$. These conditional probabilities were estimated as $\Pr(p, n_2|x) = \#(x, p, n_2)/\#(x)$, where x can be n_1 or v. Using inflected forms, he was able to classify 85% of his examples with 75% accuracy.

Calvo & Gelbukh (2003) used exact phrases instead of the `NEAR` operator, and compared the frequencies "$v\ p\ n_2$" and "$n_1\ p\ n_2$". For example, to disambiguate *Veo al gato con un telescopio*, they compared the n-gram frequencies for phrases "*veo con telescopio*" and "*gato con telescopio*". They tested this idea on 181 randomly chosen Spanish examples, and were able to classify 89.5% of them with a accuracy of 91.97%.

Lapata & Keller (2005) used exact-phrase queries in *AltaVista* and the models of Volk (2000), Volk (2001) and Calvo & Gelbukh (2003). On a random subset of 1,000 examples from *Ratnaparkhi's dataset*, they achieved 69.40% accuracy (baseline 56.80%), which they found to be worse than using *BNC* with the same models: 74.40%.

6.1.3 Models and Features

As in Chapter 3, I use three kinds of features, all derived from the Web: (a) word association scores; (b) paraphrases; and (c) surface features. I used the training part of *Ratnaparkhi's dataset* when developing the features in the last two groups.

n-gram Models

I use the following co-occurrence models:

(1) $\#(n_1, p)$ vs. $\#(v, p)$
(2) $\Pr(p|n_1)$ vs. $\Pr(p|v)$
(3) $\#(n_1, p, n_2)$ vs. $\#(v, p, n_2)$
(4) $\Pr(p, n_2|n_1)$ vs. $\Pr(p, n_2|v)$

In the above models, the left score is associated with a *noun attachment*, and the right score with a *verb attachment*. I estimate the n-gram counts using inflected exact phrase queries against *MSN Search*; the conditional probabilities are estimated as ratios of n-gram frequencies as described in section 6.1.2 above. I allow for determiners where appropriate, e.g., between the preposition and the noun when querying for $\#(v, p, n_2)$, and I add up the frequencies for all possible variations of the query. I tried using the class-based smoothing technique described by (Hindle & Rooth 1993), which led to better coverage for some of the models. I also tried backing off from (3) to (1), and from (4) to (2), as well as backing off plus smoothing, but there was no improvement over smoothing alone. I found n-gram counts to be unreliable when pronouns appear in the test set as n_1 or n_2, and therefore I disabled those models in these cases. Happily, such examples can still be handled by paraphrases or surface features (see below).

Web-Derived Surface Features

Consider for example, the sentence

John opened the door with his key.

which contains a difficult verb attachment example since *doors*, *keys*, and *opening* are all semantically related. To determine if this should be a verb or a noun attachment, I look for surface cues in other contexts that could indicate which of these terms tend to associate most closely. For example, if I can find an instance with parentheses like *"opened the door (with his key)"*

it would suggest a verb attachment, since the parentheses signal that "*with a key*" acts as a unit that is independent of *door*, and therefore has to attach to the verb.

Hyphens, colons, capitalization, and other punctuation can also help making disambiguation decisions. For example, when disambiguating the sentence

John eats spaghetti with sauce.

finding the following example on the Web

"eat: spaghetti with sauce"

would suggest a noun attachment.

Table 6.1 illustrates a variety of surface features, along with the attachment decisions they are assumed to suggest (singleton events are ignored). The table shows that such surface features have low coverage.

Since search engines ignore punctuation characters, I issue exact phrase inflected queries against *Google* and then post-process the top 1,000 resulting summaries, looking for the surface features of interest.

Paraphrases

The second way I extend the use of Web counts is by paraphrasing the relation in an alternative form that could suggest the correct attachment. Given a quadruple (v, n_1, p, n_2), I look on the Web for instances of the following patterns:

Surface Feature	**Prediction**	**Accuracy (%)**	**Coverage (%)**
open Door with a key	noun	100.00	0.13
(open) door with a key	noun	66.67	0.28
open (door with a key)	noun	71.43	0.97
open - door with a key	noun	69.70	1.52
open / door with a key	noun	60.00	0.46
open, door with a key	noun	65.77	5.11
open: door with a key	noun	64.71	1.57
open; door with a key	noun	60.00	0.23
open. door with a key	noun	64.13	4.24
open? door with a key	noun	83.33	0.55
open! door with a key	noun	66.67	0.14
open door With a Key	verb	0.00	0.00
(open door) with a key	verb	50.00	0.09
open door (with a key)	verb	73.58	2.44
open door - with a key	verb	68.18	2.03
open door / with a key	verb	100.00	0.14
open door, with a key	verb	58.44	7.09
open door: with a key	verb	70.59	0.78
open door; with a key	verb	75.00	0.18
open door. with a key	verb	60.77	5.99
open door! with a key	verb	100.00	0.18
Surface features (sum)		**73.13±5.41**	**9.26**

Table 6.1: **Surface features for PP-attachment 2,171.** Accuracy and coverage (in %s) are shown across all examples, and are illustrated by the quadruple (*open*, *door*, *with*, *key*).

(1)	$v\ n_2\ n_1$	(noun)
(2)	$v\ p\ n_2\ n_1$	(verb)
(3)	$p\ n_2\ *\ v\ n_1$	(verb)
(4)	$n_1\ p\ n_2\ v$	(noun)
(5)	v *pronoun* $p\ n_2$	(verb)
(6)	*be* $n_1\ p\ n_2$	(noun)

These patterns are linguistically motivated and each one supports either a noun or a verb attachment, as indicated above.

Pattern (1) predicts a noun attachment if "$n_1\ p\ n_2$" can be expressed as a noun compound "$n_2\ n_1$", which would mean that the verb has a single object. For example, pattern (1) would paraphrase

$$\textit{meet/v demands/}n_1\ \underline{\textit{from/p customers/}n_2}$$

as

$$\textit{meet/v the}\ \underline{\textit{customer/}n_2}\ \textit{demands/}n_1$$

Note that in case of ditransitive verbs, the pattern could make an incorrect prediction. For example

$$\textit{give/v an apple/}n_1\ \underline{\textit{to/p the boy/}n_2} \rightarrow \textit{give/v}\ \underline{\textit{the boy/}n_2}\ \textit{an apple/}n_1$$

where "*boy an apple*" clearly is not a noun compound, and the verb still has two objects: only the order of the direct and the indirect object has changed.

In order to prevent this, in the paraphrase, I do not allow a determiner before n_1, and I do require one before n_2. In addition, I disallow the pattern if the preposition is *to* and I require both n_1 and n_2 to be nouns (as opposed to numbers, percents, pronouns, determiners, etc.).

Pattern (2) predicts a verb attachment. It presupposes that "p n_2" is an indirect object of the verb v and tries to switch it with the direct object n_1, which is possible in some cases, e.g.,

had/v a program/n$_1$ <u>*in/p place/n*$_2$</u>

would be transformed into

had/v <u>*in/p place/n*$_2$</u> *a program/n*$_1$

I require that n_1 be preceded by a determiner in order to prevent "n_2 n_1" from forming a noun compound in the paraphrase.

Pattern (3) predicts a verb attachment. It looks for appositions, where the prepositional phrase "p n_2" has moved in front of the verb, e.g.,

I gave/v an apple/n$_1$ <u>*to/p the boy/n*$_2$</u>.

can be paraphrased as

It was <u>*to the boy/n*$_2$</u> *that I gave/v an apple/n*$_1$.

I allow zero or more (up to three) intervening words at the position of the asterisk in pattern (3): "p n_2 * v n_1".

Pattern (4) predict a noun attachment. It looks for appositions, where the whole complex NP "n_1 p n_2" has moved in front of the verb v. For example, it would transform

shaken/v <u>*confidence/n*$_1$ *in/p markets/n*$_2$</u>

into

$$\underline{\textit{confidence}/n_1\ \textit{in}/p\ \textit{markets}/n_2}\ \textit{shaken}/v$$

Pattern (5) is motivated by the observation that prepositional phrases do not like attaching to pronouns (Hindle & Rooth 1993). Therefore, if n_1 is a pronoun, the probability of a verb attachment is very high. (I have a separate model that checks whether n_1 is a pronoun.) Pattern (5) substitutes n_1 with a dative pronoun: *him* or *her*. For example, it would paraphrase

$$\textit{put}/v\ \underline{\textit{a client}/n_1}\ \textit{at}/p\ \textit{odds}/n_2$$

as

$$\textit{put}/v\ \underline{\textit{him}}\ \textit{at}/p\ \textit{odds}/n_2$$

Pattern (6) is motivated by the observation that the verb *to be* is typically used with a noun attachment. (I have a separate model that checks whether v is a form of the verb *to be*.) The pattern substitutes v with *is* and *are*, e.g., it will turn

$$\textit{eat}/v\ \textit{spaghetti}/n_1\ \textit{with}/p\ \textit{sauce}/n_2$$

into

$$\textit{is}\ \textit{spaghetti}/n_1\ \textit{with}/p\ \textit{sauce}/n_2.$$

All six patterns allow for determiners where appropriate, unless otherwise stated. A prediction is made if at least one instance of the pattern has been found.

6.1.4 Experiments and Evaluation

In order to make my results directly comparable to those of other researchers, I perform the evaluation on the test part of *Ratnaparkhi's dataset*, consisting of 3,097 quadruples (v, n_1, p, n_2). However, there are some problems with that dataset, due primarily to extraction errors. First, the dataset contains 149 examples in which a bare determiner like *the*, *a*, *an*, etc., is labeled as n_1 or n_2, instead of the actual head noun, e.g., (*is*, *<u>the</u>*, *of*, *kind*), (*left*, *chairmanship*, *of*, *<u>the</u>*), (*acquire*, *securities*, *for*, *<u>an</u>*), etc. While supervised algorithms can compensate for the problem by learning from the training set that *the* can be a "noun", unsupervised algorithms cannot do so. Second, there are 230 examples in which the nouns contain symbols like %, /, &, ', which cannot be used in queries against a search engine, e.g., (*buy*, *<u>%</u>*, *for*, *10*), (*beat*, *<u>S&P-down</u>*, *from*, *<u>%</u>*), (*is*, *<u>43%-owned</u>*, *by*, *firm*), etc. While being problematic for my models, such quadruples do not necessarily represent errors of extraction when the dataset was created.

Following Ratnaparkhi (1998) and Pantel & Lin (2000), I predict a noun attachment for all 926 test examples whose preposition is *of* (which is very accurate: 99.14%), and I only try to classify the remaining 2,171 test examples, which I will call *Ratnaparkhi's dataset*-2,171. The performance of the individual models on that smaller dataset is shown in Table 6.2. The table also shows 95%-level confidence intervals for the accuracy, calculated as described in section 3.9.4. The baseline accuracy of always assigning verb attachment is 58.18%, which is close to the accuracy for three of the models: "$\#(v, p) : \#(n_1, p)$" with 58.91%, "$v\ n_2\ n_1$" with 59.29%, and "$p\ n_2\ v\ n_1$" with 57.79%.

The poor performance of these models probably stems to an extent from the

Model	**Accuracy (%)**	**Coverage (%)**
Baseline-2,171 (verb attachment)	58.18±2.09	100.00
$\#(v,p) : \#(n_1,p)$	58.91±2.27	83.97
$\Pr(p\|v) : \Pr(p\|n_1)$	66.81±2.19	83.97
$\Pr(p\|v) : \Pr(p\|n_1)$ smoothed	**66.81±2.19**	83.97
$\#(v,p,n_2) : \#(n_1,p,n_2)$	65.78±2.25	81.02
$\Pr(p,n_2\|v) : \Pr(p,n_2\|n_1)$	68.34±2.20	81.62
$\Pr(p,n_2\|v) : \Pr(p,n_2\|n_1)$ smoothed	**68.46±2.17**	83.97
(1) "$v\ n_2\ n_1$"	**59.29±4.46**	22.06
(2) "$p\ n_2\ v\ n_1$"	**57.79±2.47**	71.58
(3) "n_1 * $p\ n_2\ v$"	**65.78±4.50**	20.73
(4) "$v\ p\ n_2\ n_1$"	**81.05±6.17**	8.75
(5) "v *pronoun* $p\ n_2$"	**75.30±3.43**	30.40
(6) "*be* $n_1\ p\ n_2$"	**63.65±3.73**	30.54
n_1 is a *pronoun*	**98.48±6.58**	3.04
v is the verb *to be*	**79.23±6.03**	9.53
Surface features	**73.13±6.52**	9.26
of → noun, majority vote	85.01±1.36	91.77
of → noun, majority vote, verb	83.63±1.34	100.00
of → noun, bootstrapping, maj. vote	84.91±1.33	96.93
of → noun, bootstrapping, maj. vote, verb	84.37±1.32	100.00

Table 6.2: **PP-attachment: evaluation results.** The results for the baseline and for the individual models are calculated on *Ratnaparkhi's dataset*-2,171; the last four lines show the performance of majority vote combinations on the full *Ratnaparkhi's dataset*.

Model	**Accuracy (%)**
Baseline 1: *noun attachment*	58.96±1.74
Baseline 2: *of* → noun; the rest → *verb attachment*	70.42±1.63
Baseline 3: *most likely for each p*	72.20±1.60
(Lapata & Keller 2005) – *AltaVista*	69.40±2.93
(Lapata & Keller 2005) – *BNC*	74.40±2.79
(Ratnaparkhi 1998)	81.60±1.40
(Pantel & Lin 2000)	84.31±1.32
My bootstrapped majority vote classifier	**84.37±1.32**
Average human: *quadruple*	88.20
Average human: *whole sentence*	93.20

Table 6.3: **PP-attachment: comparisons on *Ratnaparkhi's dataset*.** Shown are human performance, different baselines, and previous unsupervised results. Note that the results of Lapata & Keller (2005) are calculated for 1,000 random examples from *Ratnaparkhi's dataset* (total: 3,097 testing examples).

removal of the quadruples whose preposition is *of*. For example, they are likely to be paraphrasable as noun compounds (e.g., *includes refinancing of debt* $\rightarrow$ *includes debt refinancing*), and therefore many of them would have been classified correctly by paraphrasing pattern (1). Most of the n-gram based word association models cover about 81-83% of the examples with about 66-68% accuracy. They are outperformed by two of the paraphrasing patterns, which show significantly better accuracy, but much lower coverage: (4) "$v\ p\ n_2\ n_1$" (81.05 accuracy, 8.75% coverage) and (5) "v *pronoun* $p\ n_2$" (75.30% accuracy, 30.40% coverage). Checking if n_1 is a pronoun yields 98.48% accuracy, but has a very low coverage: 3.04% Checking if v is a form of the verb *to be* has a better coverage, 9.53%, and a very good accuracy: 79.23%. The surface features are represented in the table by a single model: for a given example, I add together the number of noun attachment predicting matches and I compare that number to the total number of verb attachment predicting matches. This yields 73.13% accuracy and 9.26% coverage. The performance for the individual surface features is shown in Table 6.1. Some of them have very good accuracy, but also very low coverage; this is why I add them together.

The last four lines of Table 6.2 show the performance of some majority vote combinations on the full *Ratnaparkhi's dataset*. I first assign the 926 *of*-examples to *noun attachment*, and then, for the remaining 2,171 examples, I use a majority vote which combines the bold rows in Table 6.2. This yields 85.01% accuracy, and 91.77% coverage. In order to achieve 100% coverage, I further assign all undecided examples to *verb attachment*, which gives 83.63% accuracy.

The last two lines in Table 6.2 show the results of bootstrapping the above-described majority voting algorithm, which made attachment predictions for 1,992 of the 2,171 examples. I use these examples and the voting predictions as training data for an n-gram based back-off classifier, similar to the one used by Collins & Brooks (1995), but limited to words and bigrams only.

First, following Collins & Brooks (1995), I normalize the examples by substituting all 4-digit numbers (possibly followed by '*s*', e.g., *1990s*) with `YEAR`, all other numbers and %s with `NUM`, all pronouns with `PRO`, the articles *a*, *an* and *the* with `ART`, and all other determiners (e.g., *this*, *one*) with `DET`. I also lemmatize all nouns and verbs using *WordNet*.

Then I use the following ratio

$$R_1 = \frac{\#(v, p|noun) + \#(n1, p|noun) + \#(p, n_2|noun)}{\#(v, p) + \#(n1, p) + \#(p, n_2)} \tag{6.1}$$

where the counts are estimated from the 1,992 training examples (not from the Web, which makes the higher-order n-grams unreliable); conditioning on *noun* means that the count is calculated over the examples that had a *noun attachment* predicted.

I choose a *noun attachment* if $R_1 > 0.5$, and a *verb attachment* if $R_1 < 0.5$. I only make a decision if the denominator is greater than 3. If no decision can be made, I back-off to the following ratio:

$$R_2 = \frac{\#(p|N)}{\#(p)} \tag{6.2}$$

Finally, I add this new model as an additional voter in the majority vote, and I obtain 84.91% accuracy, and 96.93% coverage. The last line in Table 6.2 shows the results when I further assign all undecided examples to *verb attachment*, which yields 84.37% accuracy.

As Table 6.3 shows, my latest result (84.37% accuracy) is as strong as that of the best unsupervised approach on this collection: Pantel & Lin (2000) achieved 84.31%. Unlike their work, I do not need a collocation database, a thesaurus, a dependency parser, nor a large domain-dependent text corpus, which makes my approach easier to implement and to extend to other languages.

Table 6.3 shows the results of other previous researchers on *Ratnaparkhi's dataset*. It also shows the accuracy for three different baselines:

- Always predicting a *noun attachment*;
- Predicting *noun attachment* if the preposition is *of*, and a *verb attachment* otherwise;
- Make the most likely prediction for each preposition.

The last two baselines are pretty high, accuracy in the low seventies, which is higher than the *AltaVista* results of Lapata & Keller (2005). However, my results and those of Pantel & Lin (2000) are statistically better than the all three baselines and than all other previously published results.

6.1.5 Conclusion and Future Work

I have shown that simple unsupervised models that make use of n-grams, surface features and paraphrases extracted from the largest existing corpus, the Web, are effective for solving the problem of prepositional phrase attachment, yielding 84.37% accuracy, which matches the best previously published unsupervised results. The approach does not require labeled training data, lexicons, or ontologies, thesaurus, parsers, which makes it promising for a wide range of other NLP tasks.

There are many ways in which the presented approach can be extended. First, there should be many more linguistically-motivated paraphrases and surface features that are worth exploring. For example, Hindle & Rooth (1993) mention some interesting heuristics. They predict a *verb attachment* if the verb is passive (unless the preposition is *by*). They also mention that if the NP object includes a superlative adjective as a pre-modifier then the noun attachment is certain, which I could use as a paraphrase pattern. In addition, when annotating the testing data, they always attach *light verbs* to the noun, and *small clauses* to the verb. Many other interesting features proposed and used by Olteanu & Moldovan (2005) are worth trying as well.

6.2 Noun Phrase Coordination

6.2.1 Introduction

Coordinating conjunctions, such as *and*, *or*, *but*, etc., pose major challenges to parsers; their proper handling is also essential for understanding the semantics of the sentence. Consider the following "cooked" example:

> "*The Department of Chronic Diseases* ***and*** *Health Promotion leads* ***and*** *strengthens global efforts to prevent* ***and*** *control chronic diseases* ***or*** *disabilities* ***and*** *to promote health* ***and*** *quality of life.*"

Coordinating conjunctions can link two words, two constituents (e.g., NPs), two clauses, two sentences, etc. Therefore, the first challenge is to identify the boundaries of the conjuncts of each coordination. The next problem comes from the interaction of the coordinations with other constituents that attach to its conjuncts (most often prepositional phrases). In the example above, we need to decide between the following two bracketings: [*health and* [*quality of life*]] and [[*health and quality*] *of life*]. From a semantic point of view, we need to determine whether the *or* in *chronic diseases or disabilities* really means *or* or is used as an *and* (Agarwal & Boggess 1992). Finally, there is an ambiguity between *NP-coordination* and *noun-coordination*, i.e. between [[*chronic diseases*] *or* [*disabilities*]] and [*chronic* [*diseases or disabilities*]].

Below I focus on a special case of the latter problem. Consider the noun phrase *car and truck production*. Its actual meaning is *car production and truck production*. However, for reasons of economy of expression, the first instance of *production* has been left out. By contrast, in *president and chief executive*, *president* is coordinated with *chief executive*, and nothing has been compressed out (it does not mean *president executive and chief executive*).

There is also a third option, an all-way coordination, where the coordinated parts are inseparable from the whole, as in *Securities and Exchange Commission*.

More formally, I consider quadruples (n_1, c, n_2, h), where n_1 and n_2 are nouns, c is a coordinating conjunction, and h is the head noun[1]. I further limit c to be *and* or *or* only. The task is to decide between an NP-coordination and a noun-coordination, given the quadruple only and independently of the local context. Syntactically, the distinction can be expressed by the following bracketings:

$$\begin{array}{ll} [\ n_1\ c\ n_2\ ]\ h & \text{(noun coordination)} \\ [\ n_1\ ]\ c\ [\ n_2\ h\ ] & \text{(NP coordination)} \end{array}$$

In order to make the task more realistic (from a parser's perspective), I ignore the option of all-way coordination and I try to predict the bracketing in *Penn Treebank* (Marcus *et al.* 1994) for configurations of this kind The *Penn Treebank* has a flat NP in case of noun-coordination, e.g.,

```
(NP car/NN and/CC truck/NN production/NN)
```

In case of NP-coordination, there is an NP that contains two internal NPs:

```
(NP
    (NP president/NN)
    and/CC
    (NP chief/NN executive/NN))
```

All-way coordinations can appear bracketed either way and make the task harder.

[1]Quadruples of the kind (n, h_1, c, h_2), e.g. *company/n cars/*h_1 *and/c trucks/*h_2, can be handled in a similar way.

6.2.2 Related Work

Coordination ambiguity is under-explored, despite being one of the major sources of structural ambiguity (e.g., *and* and *or* account for about 3% of the word tokens in *BNC*), and despite being the hardest type of dependency to parse (e.g., Collins (2003) reports only 61.47% recall and 62.20% precision when parsing dependencies involving coordination).

Rus *et al.* (2002) present a deterministic rule-based approach for bracketing *in context* of coordinated NPs of the kind "n_1 c n_2 h" as a necessary step towards logical form derivation. Their algorithm uses POS tagging, syntactic parsing, manually annotated semantic senses, lookups in a semantic network (*WordNet*), and the type of the coordination conjunction (e.g., *and*, *or*), in order to make a three-way decision: NP-coordination, noun-coordination, and all-way coordination. Using a back-off sequence of three different heuristics, they achieve 83.52% accuracy (baseline 61.52%) on a set of 298 examples. When three additional context-dependent heuristics and 224 additional examples with local contexts were added, the accuracy jumped to 87.42% (baseline 52.35%), with 71.05% coverage.

Resnik (1999b) tries to bracket the following kinds of three- and four-noun configurations: "n_1 *and* n_2 n_3" and "n_1 n_2 *and* n_3 n_4". While there are two bracketing options for the former, five valid bracketings exist for the latter. Following Kurohashi & Nagao (1992), Resnik makes decisions based on similarity of form (i.e., number agreement: 53% accuracy, 90.6% coverage), similarity of meaning (66% accuracy, 71.2% coverage), and conceptual association (75% accuracy, 69.3% coverage). Using a decision tree , he achieves 80% accuracy (baseline 66%) at 100% coverage for the three-noun coordinations. For the four-noun coordinations the accuracy is 81.6% (baseline 44.9%), 85.4% coverage.

Chantree *et al.* (2005) cover a large set of bracketing ambiguities, not limited to nouns. Using distributional information from the *BNC*, they calculate similarities between words, similarly to Resnik (1999b), and achieve an F-measure below 50%.

Goldberg (1999) brackets phrases of the kind "n_1 p n_2 c n_3", e.g., *box/*n_1 *of/p chocolates/*n_2 *and/c roses/*n_3. Using an adaptation of the maximum entropy algorithm used by Ratnaparkhi (1998) for PP-attachment, she achieves 72% accuracy (baseline 64%).

Agarwal & Boggess (1992) focus on determining the boundaries of the conjuncts of coordinate conjunctions. Using POS and case labels in a deterministic algorithm, they achieve 81.6% accuracy. Kurohashi & Nagao (1992) work on the same problem for Japanese. Their algorithm looks for similar word sequences among with sentence simplification and achieves 81.3% accuracy. Okumura & Muraki (1994) address that problem heuristically using orthographical, syntactic, and semantic information.

In recent work, Hogan (2007) presents a method for improving NP coordination disambiguation within the framework of a lexicalised history-based parsing model. Using two main information sources, symmetry in conjunct structure and dependency between conjunct's lexical heads, he achieves 73.8% F-measure (baseline: 69.9%). This work is related to that of Ratnaparkhi *et al.* (1994a) and Charniak & Johnson (2005), who use specialized features targeting coordination in discriminative re-rankers for parsing.

Buyko *et al.* (2007) resolve coordination ellipses for biological named entities. Using conditional random fields (CRFs), they achieve 93% F-measure on the GENIA corpus. On the same corpus, Shimbo & Hara (2007) apply a sequence-alignment model for detecting and disambiguating coordinate conjunctions, achieving F-measure of 70.5% and 57.2%, respectively.

6.2.3 Models and Features

n-gram Models

I use the following n-gram models:

(*i*) $\#(n_1, h)$ vs. $\#(n_2, h)$

(*ii*) $\#(n_1, h)$ vs. $\#(n_1, c, n_2)$

Model (*i*) compares how likely it is for n_1 to modify h, as opposed to n_2 modifying h. Model (*ii*) checks which association is stronger: between n_1 and h, or between n_1 and n_2. When calculating the frequency for $\#(n_1, c, n_2)$, I query for both *or* and *and* and I add up the counts, regardless of whether the original coordination c was *or* or *and*.

Paraphrasing Patterns

I use the following paraphrasing patterns:

(1)	$n_2\ c\ n_1\ h$	(noun-coordination)
(2)	$n_2\ h\ c\ n_1$	(NP-coordination)
(3)	$n_1\ h\ c\ n_2\ h$	(noun-coordination)
(4)	$n_2\ h\ c\ n_1\ h$	(noun-coordination)

If matched frequently enough, the patterns predict the coordination decision indicated in parentheses. If not found or found infrequently, the opposite decision is made.

Pattern (1) predicts a noun-coordination. It switches the places of n_1 and n_2 in the coordinated NP, e.g.,

$$\underline{bar/n_1}\ and/c\ \boldsymbol{pie/}n_2\ graph/h$$

is paraphrased as

pie/n_2 *and/c* *bar/*n_1 *graph/h*

Pattern (2) predicts an NP-coordination. It moves n_2 and h together to the left of the coordination conjunction, and places n_1 to the right, e.g.,

president/n_1 *and/c* ***chief/***n_2 ***executive/****h*

is paraphrased as

chief/n_2 ***executive/****h* *and/c* *president/*n_1

Pattern (3) predicts a noun-coordination. It inserts the elided head h after n_1 with the hope that if there is ellipsis of the head h, a coordination involving the full phrase "n_1 h" will be likely to be found elsewhere on the Web, e.g.,

bar/n_1 *and/c* *pie/*n_2 *graph/h*

is paraphrased as

bar/n_1 *graph/h* *and/c* *pie/*n_2 *graph/h*

Pattern (4) predicts a noun-coordination. It combines pattern (1) and pattern (3) by not only inserting h after n_1, but also switching the places of n_1 and n_2. For example,

bar/n_1 *and/c* *pie/*n_2 *graph/h*

is paraphrased as

pie/n_2 *graph/h* *and/c* ***bar/***n_2 ***graph/****h*

Heuristics

I also included some of the heuristics proposed by Rus *et al.* (2002), as shown in Table 6.5. Heuristic 1 predicts an NP-coordination when n_1 and n_2 are the same word, e.g., *milk/*n_1 *and/c milk/*n_2 *products/h*. Heuristics 2 and 3 perform a lookup in *WordNet* and I decided not not to use them. Heuristics 4, 5 and 6 exploit the local context, namely the adjectives modifying n_1 and/or n_2. Heuristic 4 predicts an NP-coordination if both n_1 and n_2 are modified by adjectives. Heuristic 5 predicts a noun-coordination if c is *or* and n_1 is modified by an adjective, but n_2 is not. Heuristic 6 predicts NP-coordination if n_1 is not modified by an adjective, but n_2 is.

Since I target coordination disambiguation independent of context, and since search engines lack POS annotations, I could not use the original heuristics 4, 5 and 6. Therefore, in my experiments, I used adapted versions of these heuristics that look for a determiner (e.g., *a, an, the*) rather than an adjective.

Number Agreement

I also include the number agreement model proposed by Resnik (1993), which makes prediction as follows:

(a) if n_1 and n_2 match in number, but n_1 and h do not, predict noun-coordination;

(b) if n_1 and n_2 do not match in number, but n_1 and h do, predict NP-coordination;

(c) otherwise leave undecided.

Surface Feature	Prediction	Accuracy (%)	Coverage (%)
(buy) and sell orders	NP-coordination	33.33	1.40
buy (and sell orders)	NP-coordination	70.00	4.67
buy: and sell orders	NP-coordination	0.00	0.00
buy; and sell orders	NP-coordination	66.67	2.80
buy. and sell orders	NP-coordination	68.57	8.18
buy[...] and sell orders	NP-coordination	49.00	46.73
buy- and sell orders	noun-coordination	77.27	5.14
buy and sell / orders	noun-coordination	50.54	21.73
(buy and sell) orders	noun-coordination	92.31	3.04
buy and sell (orders)	noun-coordination	90.91	2.57
buy and sell, orders	noun-coordination	92.86	13.08
buy and sell: orders	noun-coordination	93.75	3.74
buy and sell; orders	noun-coordination	100.00	1.87
buy and sell. orders	noun-coordination	93.33	7.01
buy and sell[...] orders	noun-coordination	85.19	18.93
Surface features (sum)		**82.80±8.93**	**21.73**

Table 6.4: **NP coordination surface features.** Accuracy and coverage shown are across all examples, not just the *buy and sell orders* shown.

Web-Derived Surface Features

The set of surface features is similar to the one I used for PP-attachment in section 6.1, and include brackets, slash, comma, colon, semicolon, dot, question mark, exclamation mark, and any special character. There are two additional NP-coordination predicting features: a dash after n_1 and a slash after n_2, see Table 6.4.

6.2.4 Evaluation

I evaluate the above-described models on a collection of 428 quadruples of the form (n_1, c, n_2, h), extracted from the *Penn Treebank*. I look for NPs of the following kinds:

$$(\text{NP} \ldots n_1/\text{N}\ c/\text{CC}\ n_2/\text{N}\ h/\text{N})$$

and

Model	Accuracy (%)	Coverage (%)
Baseline: *noun-coordination*	56.54±4.73	100.00
(n_1, h) vs. (n_2, h)	**80.33±7.93**	28.50
(n_1, h) vs. (n_1, c, n_2)	61.14±7.03	45.09
(n_2, c, n_1, h)	**88.33±10.51**	14.02
(n_2, h, c, n_1)	**76.60±9.50**	21.96
(n_1, h, c, n_2, h)	**75.00±18.36**	6.54
(n_2, h, c, n_1, h)	**78.67±10.55**	17.52
Heuristic 1	**75.00±44.94**	0.93
Heuristic 4	64.29±18.46	6.54
Heuristic 5	61.54±13.58	12.15
Heuristic 6	**87.09±15.95**	7.24
Number agreement	**72.22±6.62**	46.26
Surface features (sum)	**82.80±8.93**	21.73
Majority vote	83.82±4.25	80.84
Majority vote, N/A → no ellipsis	***80.61±4.01***	***100.00***

Table 6.5: **NP Coordination: evaluation results.**

(NP (NP ... n_1/N) c/CC (NP ... n_2/N h/N))

The nouns n_1, n_2 and h can be POS tagged as NN, NNS, NNP, or NNPS, and the NPs can contain determiners and non-noun modifiers preceding n_1 and/or n_2, but no other nouns.

The full dataset is shown in Appendix D. It contains some repetitions of quadruples, e.g., the NP-coordinated (*test*, *and*, *Learning*, *Materials*); in some cases, the different instances of the same quadruple can have different labels, e.g., (*FTC*, *and*, *Justice*, *Department*) appears both noun-coordinated and NP-coordinated. Such inconsistencies stem from either context dependence or from inconsistencies in the *Penn Treebank* annotations.

As Table 6.5 shows, the n-gram model (i) performs very well (80.33% accuracy, 28.50% coverage), but the n-gram model (ii) is not that accurate (61.14% accuracy, 7.03% coverage). This is probably because the "n_1 c n_2" is a trigram, which makes it less likely to be observed than the alternative "n_1 h", which is a bigram.

Interestingly, the number agreement feature yields 72.22% accuracy and 46.26% coverage, while in the experiments of Resnik (1993) it was the best feature 90% accuracy and 53% coverage. This might indicate that my dataset is harder.

The surface features are represented in Table 6.5 by a single model: for a given example, I add up together the number of noun-coordination predicting matches and I compare that number to the total number of NP-coordination predicting matches, which yields 82.80% accuracy and 21.73% coverage. The performance for the individual surface features is shown in Table 6.4. We can see that they are very good predictors of noun-coordination, but are less reliable when predicting NP-coordination.

I combine the bold rows of Table 6.5 in a majority vote, obtaining 83.82% accuracy and 80.84% coverage. Assigning all undecided cases to NP-coordination, yields 80.61% accuracy (and 100% coverage).

6.2.5 Conclusion and Future Work

I have shown that the simple Web-based unsupervised models similar to the ones I proposed for *noun compound bracketing* in chapter 3 – n-grams, surface features and linguistically-motivated paraphrases – are effective for solving the problem of *NP coordination*: I have achieved 80.61% accuracy, which is on par with other approaches, whose best scores fall into the low 80's for accuracy; direct comparison is not possible though, as the tasks and the datasets all differ.

Other kinds of coordination-related attachment ambiguity problems might be addressable in a similar way, e.g., identification of the boundaries of the coordination conjuncts, or interactions between coordinations and prepositional phrases, with the ultimate goal of extending and applying the proposed approach to real NLP tasks, e.g., named entity recognition and syntactic parsing, as done in Buyko *et al.* (2007) and Hogan (2007), respectively.

Chapter 7

Quality and Stability of Page Hit Estimates

Web search engines provide an easy access for NLP researchers to world's biggest corpus, but not without drawbacks. In this chapter, I point to some problems and limitations of using search engine page hits as a proxy for n-gram frequency estimates. I further describe a study on the quality and stability of such estimates across search engines and over time, as well as on the impact of using word inflections and of limiting the queries to English pages. Using the task of noun compound bracketing and 14 different n-gram based models, I illustrate that while sometimes causing sizable fluctuations, variability's impact generally is not statistically significant.

An initial version of this study appeared in (Nakov & Hearst 2005b).

7.1 Using Web Page Hits: Problems and Limitations

7.1.1 Lack of Linguistic Restrictions on the Query

Many problems with page hit estimates stem from Web search engines not allowing for linguistic restrictions on the query. While some linguistic annotations might be supported and used by Web search engines internally, they are not accessible to the end user, which can limit search engine's value as a tool for corpus linguistics research.

For example, when bracketing the three-word noun compound *home health care*, the probabilistic association models described in section 3.4 would need to calculate the probability that *health* modifies *care*, i.e. $\Pr(health \rightarrow care|care) = \frac{\#(''health\ care'')}{\#(care)}$. This calculation requires estimates for the frequencies of *"health care"* and *care*, ideally where both words are used as nouns. While the exact phrase query "*health care*" would almost guarantee that both words are used nouns, a query for *care* would return many pages where that word is used as a verb, which would lead to an overestimated value for the denominator, and thus to an underestimated probability. On the other hand, since the word *health* can only be a noun, the estimate for $\Pr(home \rightarrow health|health) = \frac{\#(''home\ health'')}{\#(health)}$ would not be affected. Therefore, the *adjacency* model (see section 3.3.1) would compare one correct probability and one underestimated probability, which can cause potential problems when their values are close.

Further, as I have already mentioned in section 3.7, search engines do not support queries containing place-holders asking for a particular part-of-speech, e.g.,

stem cells VERB PREP DET brain

where the uppercase typed placeholders stand for a verb, a preposition and a determiner, respectively.

This makes typed placeholder queries prohibitively expensive: for example, in the above pattern, one would need to substitute every possible verb, every possible preposition and every possible determiner, and then to issue a separate exact-phrase query for each combination.

7.1.2 No Indexing for Punctuation

A further problem is caused by the fact that search engines ignore punctuation at indexing time. Consider for example the bigram *health care*. An exact-phrase query for "*health care*" does not guarantee that the words will be really adjacent: they might belong to different NPs, different sentences, even different paragraphs. For example, a search engine would find a "legitimate" exact-phrase bigram match in the following sentence:

> "*This section is for you if you care about* <u>*health,*</u> <u>*care*</u> *about fitness, like sports and above all, want to help people.*"[1]

As I have already mentioned in section 3.5, the lack of punctuation also makes impossible the direct execution of exact-phrase queries containing plural genitive markers (e.g., "*protein synthesis' inhibition*"), hyphens (e.g., "*law-enforcement officer*"), parentheses (e.g., "*bronchoalveolar lavage (BAL) fluid*"), etc. In the last example, a further potential problem is caused by the lack of distinction between letters in uppercase vs. in lowercase.

[1] From `http://www.jobmonkey.com/sports/html/health_fitness_careers_overvie.html`

7.1.3 Page Hits are Not n-grams

There are other reasons why using page hits as a proxy for n-gram frequencies can yield some counter-intuitive results. Consider the bigrams $w_1 w_4$, $w_2 w_4$ and $w_3 w_4$ and a page that contains each of them exactly once. A search engine will contribute a page count of 1 for w_4 instead of a frequency of 3; thus the number of page hits for w_4 can be smaller than that for the sum of the bigrams that contain it. Keller & Lapata (2003) describe more potential problems with page hits.

7.1.4 Rounding and Extrapolation of Page Hit Estimates

Another potentially undesirable aspect of using page hits for linguistic research is that two of the major Web search engines, *Google* and *Yahoo!* provide rounded estimates rather than exact numbers. For example, in June 2005, *MSN Search* returns 60,$\underline{219{,}609}$ page hits for *cell*, while *Google* and *Yahoo!* return 397,$\underline{000{,}000}$ and 637,$\underline{000{,}000}$, respectively. This rounding can be problematic, especially when comparing two ratios of page hits, e.g., $\frac{\#(stem,cell)}{\#(cell)}$ and $\frac{\#(brain,stem)}{\#(stem)}$.

The rounding is probably done since, once the numbers get somewhat large, exact counts are not necessary for the ordinary user, who hardly expects them from a search engine that only indexes part of the constantly changing Web anyway. From a technical viewpoint, calculating the exact page hits from multiple distributed and continually changing indexes is computationally expensive. Therefore, commercial search engines, which are optimized for quickly returning results to the user under high load, may sample from their indexes, rather than performing exact computations (Véronis 2005b). They would stop scanning their indexes once they are ready to return N results (typically 10, but no more than 100 at a time, and no more than 1,000 in total for a given query), and they would calculate the expected total number of pages by extrapolation, taking into account how many results have been produced so far, and what part of the index has been consumed in order to produce them. Another possible cause for obtaining inflated numbers is that page hit estimates come in part from anchor text of unexplored pages. In other words, search engines find links to pages that contain the indexed words, without necessarily crawling all those pages.

It is possible to observe page hits inflation by issuing an exact-phrase query that matches less than 1,000 actual pages. For example, at the time of writing, a *Google* query for "*searches engines*" returns 10,200 page hits, but if one actually asks to see them all, it turns out that there are only 340 actual pages.

7.1.5 Inconsistencies of Page Hit Estimates and Boolean Logic

The extrapolation process has caused various speculations and allegations. For example, in a series of publications on his popular blog[2], Jean Véronis suggested that many search engines inflate their page hits for marketing reasons. Since they never provide access to more than 1,000 result pages per query, the user cannot actually verify whether the listed total number of pages is correct. In 2005, after having observed over a period of time the frequencies of fifty English words drawn randomly from mid-range frequencies in a one million word corpus of English text, Véronis estimated that the true index sizes of *Google* and *MSN Search* are 60% and 75%, respectively, of the officially announced numbers (Véronis 2005c; Véronis 2005d). While he was unable to find similar inconsistencies for

[2]`http://aixtal.blogspot.com/`

Yahoo!, Véronis claims it suspiciously doubled its index in March 2005 (Véronis 2005e).

Véronis also reported problems with *Google*'s Boolean logic (which he called a "*Googlean logic*"). For example, in February 2005, a query for *Chirac and Sarkozy* produced about half of the page hits for *Chirac* alone. Even stranger things happened for queries like *x AND x*, *x OR x*, or for ones repeating the same word multiple times (Véronis 2005c). Apparently, these problems have been fixed (Véronis 2005a).

The actual implications of these allegations remain unclear. If the page hit estimates are consistent across different queries, this should not impact *ratios* of page hit estimates, and while some researchers do use operations like OR and NEAR, e.g., Volk (2001), Mihalcea & Moldovan (1999), Lapata & Keller (2004), this is not strictly necessary: it is possible to use multiple queries instead of OR, and NEAR could be emulated to some extent using the '*' operator in *Google*.

7.1.6 Instability of Page Hit Estimates

A major problem, from a research perspective, is caused by the instability of query results. Nowadays, Web search engines are too complex to run on a single machine, and the queries are served by multiple servers, which collaborate to produce the final result. In addition, the Web is dynamic, many pages change frequently, others disappear or are created for the first time, and therefore search engines need to update their indexes frequently; commercial search engines often compete on how "fresh" their indexes are. Finally, search engines constantly improve their algorithms for sampling from the index, for page hit extrapolation, for result ranking, etc. As a result, the number of page hits for a given query, as well as the query results in general, could change over time in unpredictable ways.

The indexes themselves are too big to be stored on a single machine and therefore are spread across multiple servers (Brin & Page 1998). For availability and efficiency reasons, there are multiple copies of the same part of the index, which are not always synchronized with one another since the different copies are updated at different times. As a result, if one issues the same query multiple times in rapid succession, connections to different physical machines could be opened, which could yield different results. This effect is known as search engine "dancing". With a little bit of luck, one can observe *Google* "dancing" by comparing the results of different data centers, e.g., `www.google.com`, `www2.google.com`, `www3.google.com`. Alternatively, one can try the *Google dance tool* at: `http://www.seochat.com/googledance`. More on the phenomenon can be found on the Web, e.g., at `http://dance.efactory.de` or simply by asking a search engine about "*Google dancing*".

From a research perspective, dynamics over time is highly undesirable, as it precludes the exact replicability of the results obtained using search engines. At best, one could reproduce the same initial conditions, and expect similar results. While this kind of variability is uncommon in NLP, where one typically experiments with a static frozen corpus, it is common in natural sciences, like physics, chemistry etc., where the same experiment is unlikely to always yield exactly the same outcome. In fact, even in computer science, for a fixed corpus, the exact replicability of the results may not be guaranteed, e.g., when the algorithm has elements of randomization, when the explanation in the paper is not detailed enough, or the authors had a programming bug, etc.

7.2 Noun Compound Bracketing Experiments

I believe that the best way to test the impact of rounding, possible inconsistencies and dynamics over time is to design a set of suitable experiments organized around a real NLP task. I chose *noun compound bracketing*, which can be solved using several different methods that make use of n-grams of different lengths. Below I perform series of experiments on *Lauer's dataset* (see section 3.8.1) comparing the accuracy of 14 different n-gram models from Chapter 3, across the following four dimensions: (1) *search engine*; (2) *time*; (3) *language filter*; and (4) *word inflection*.

For each model, I issue exact phrase queries within a single day. Unless otherwise stated, the queries are not inflected and no language filter is applied. I use a threshold on the module of the difference between the left- and the right-predicting n-gram frequencies: I only make a bracketing decision if the difference is at least five.

7.2.1 Variability over Time

I study the variability over time for *Google* and for *MSN Search*. I chose time snapshots at varying time intervals in order to lower the potential impact of major search engines' index changes, in case they are scheduled at fixed time intervals.

For *Google*, I collected n-gram statistics for October 22, 2007, and for four different dates in 2005: April 30, May 4, June 7, and June 11. The results for the accuracy are shown in Figure 7.1 (no language filter, no word inflections), Figure 7.2 (no language filter, using word inflections), Figure 7.5 (English pages only, no word inflections) and Figure 7.6 (English pages only, using word inflections). The variability in accuracy is low for the first

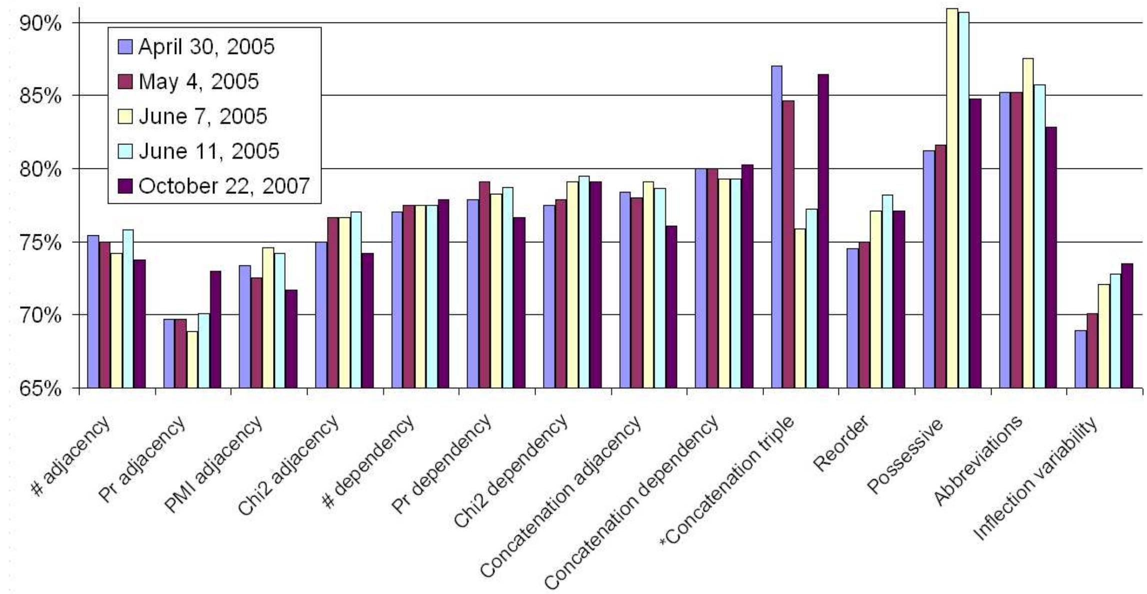

Figure 7.1: **Accuracy over time for *Google*:** pages in any language, no word inflections. The models with statistically significant variations are marked with an asterisk.

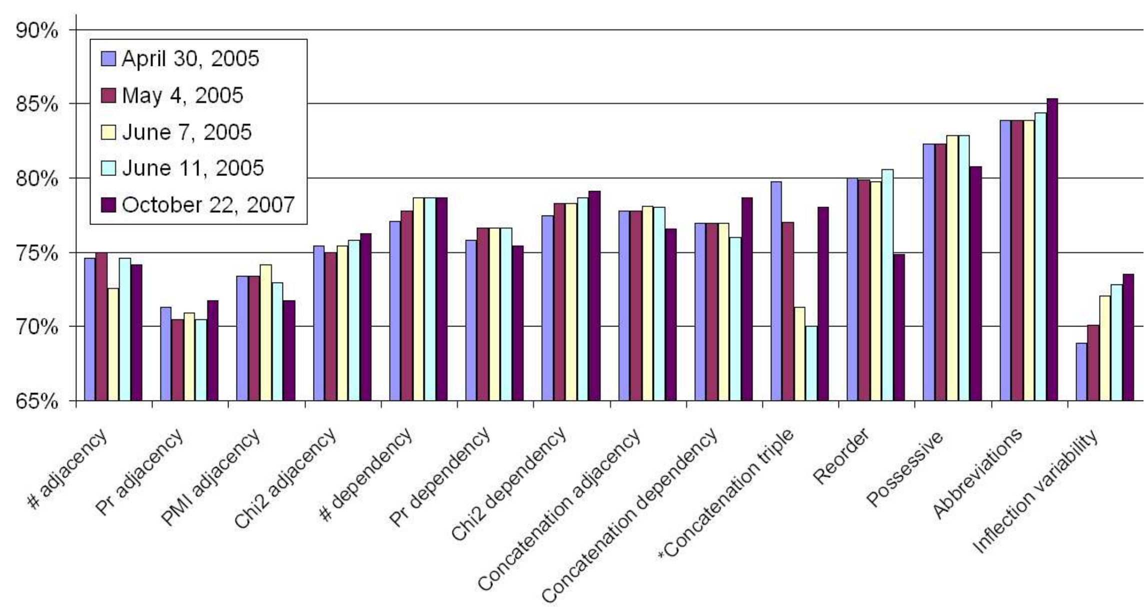

Figure 7.2: **Accuracy over time for *Google*:** pages in any language, with word inflections. The models with statistically significant variations are marked with an asterisk.

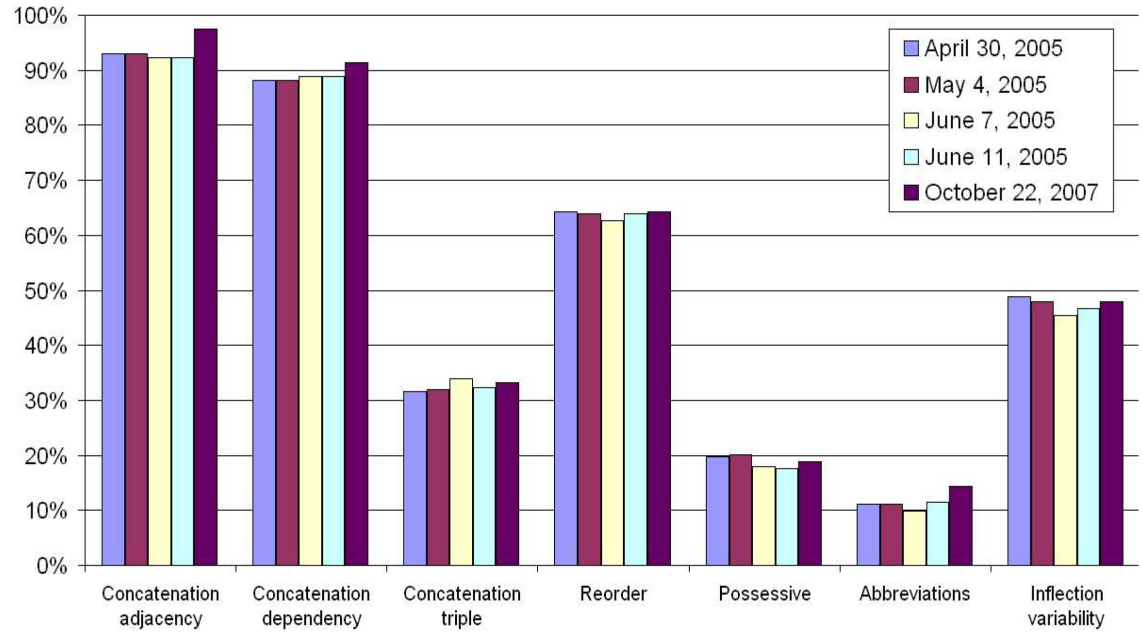

Figure 7.3: **Coverage over time for *Google*:** pages in any language, no word inflections.

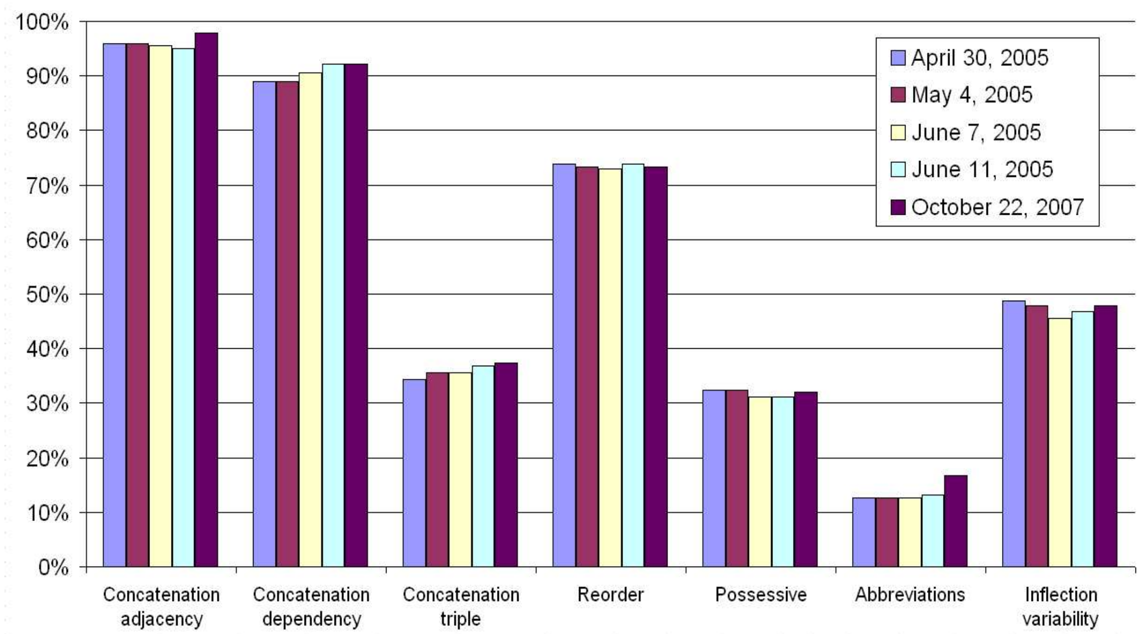

Figure 7.4: **Coverage over time for *Google*:** pages in any language, with word inflections.

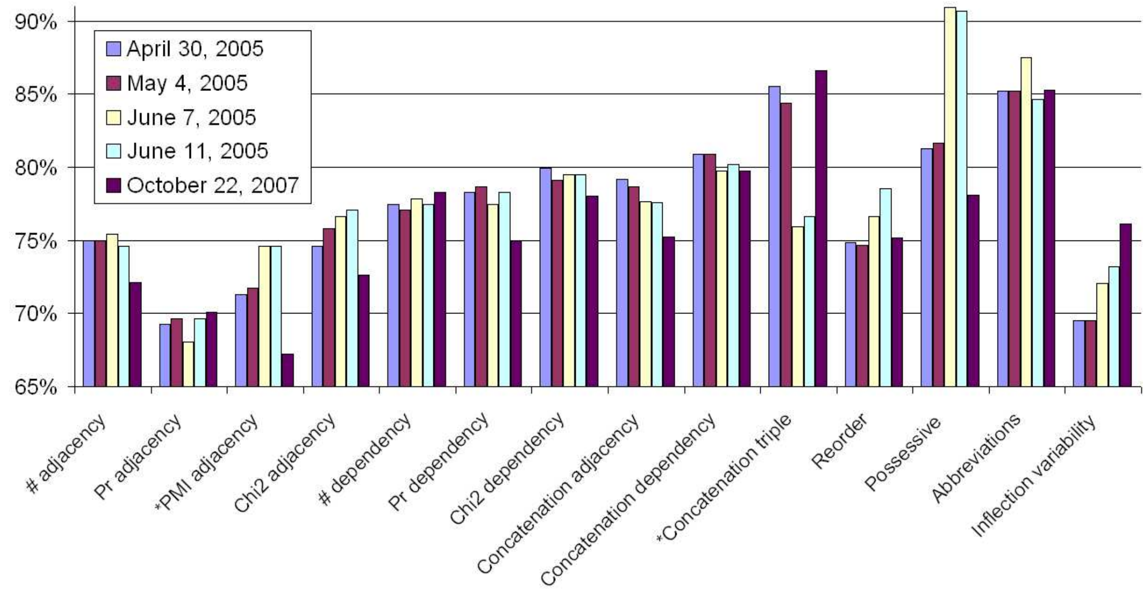

Figure 7.5: **Accuracy over time for *Google*:** English pages only, no word inflections. The models with statistically significant variations are marked with an asterisk.

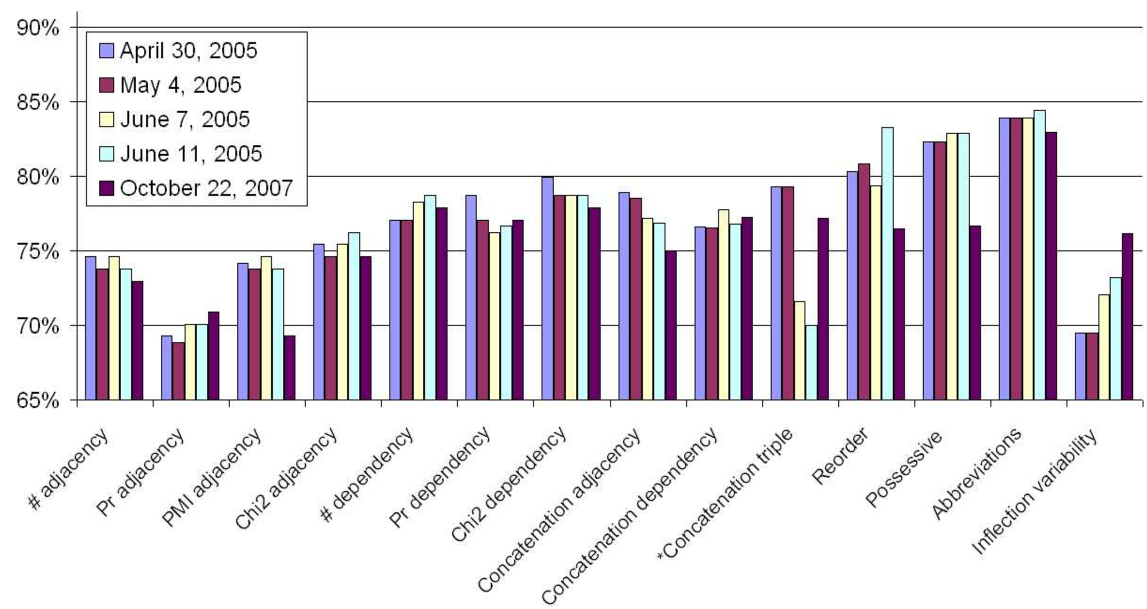

Figure 7.6: **Accuracy over time for *Google*:** English pages only, with word inflections. The models with statistically significant variations are marked with an asterisk.

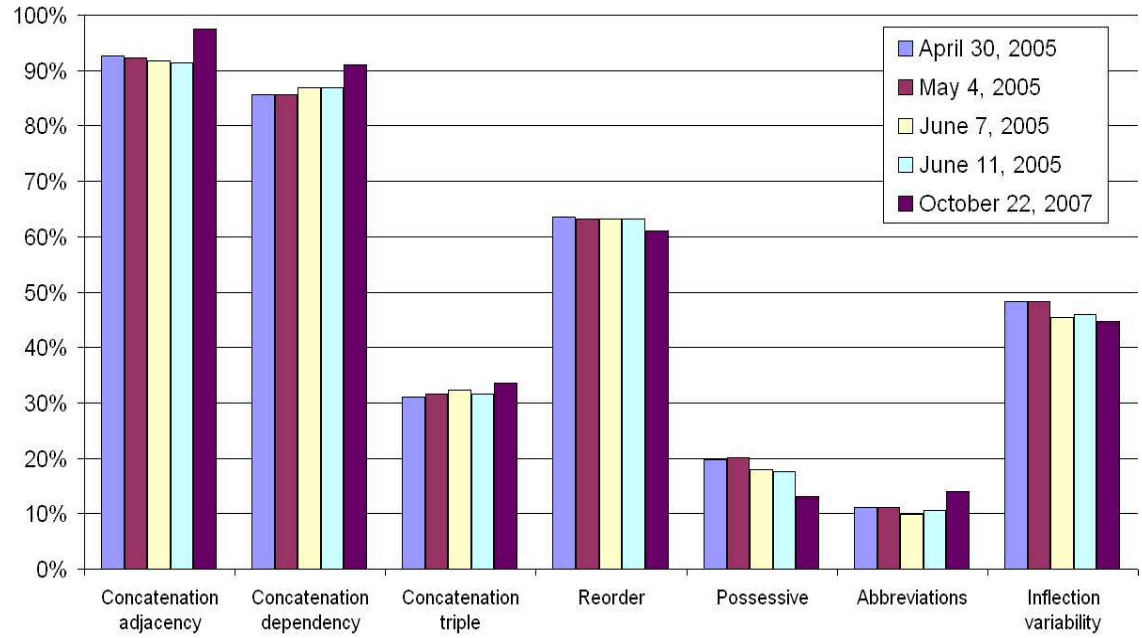

Figure 7.7: **Coverage over time for *Google*:** English pages only, no word inflections.

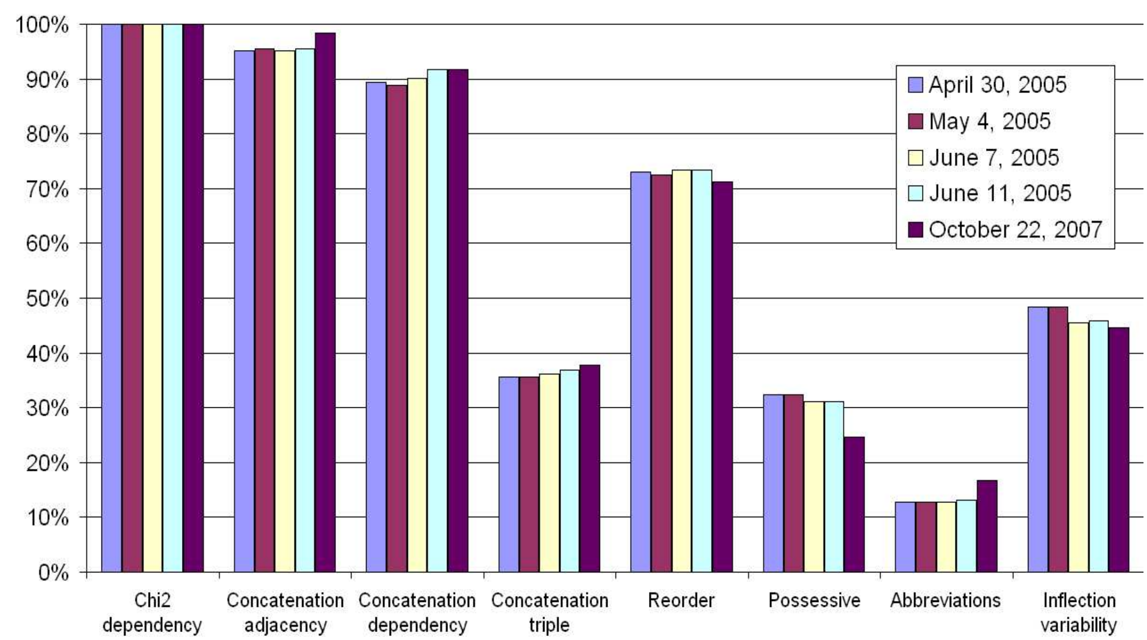

Figure 7.8: **Coverage over time for *Google*:** English pages only, with word inflections.

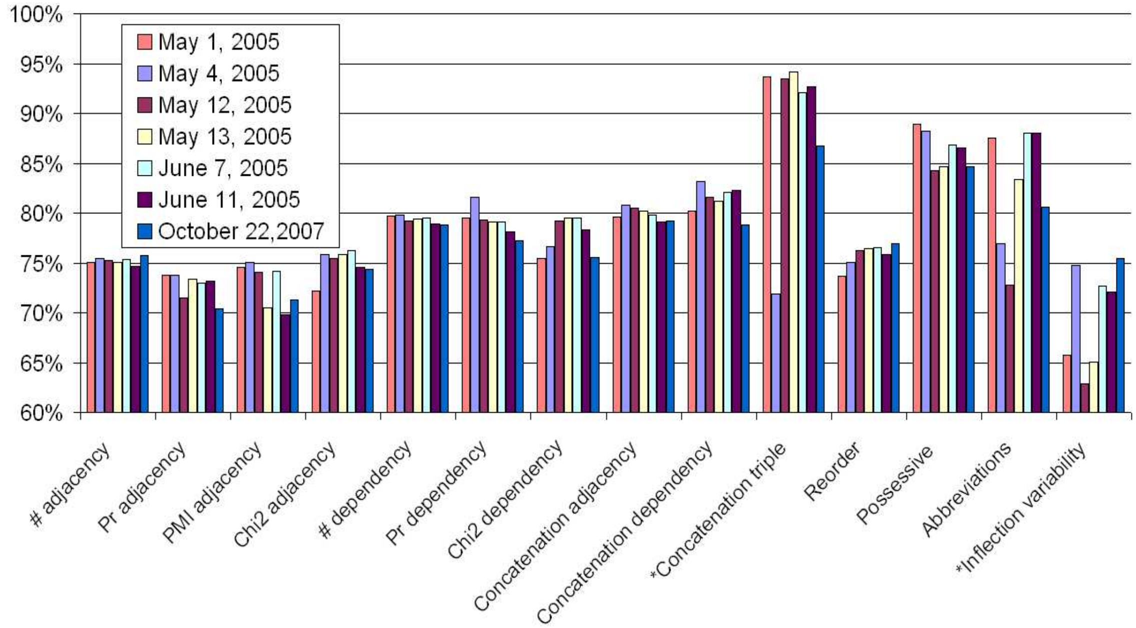

Figure 7.9: **Accuracy over time for *MSN*:** pages in any language, no word inflections. The models with statistically significant variations are marked with an asterisk.

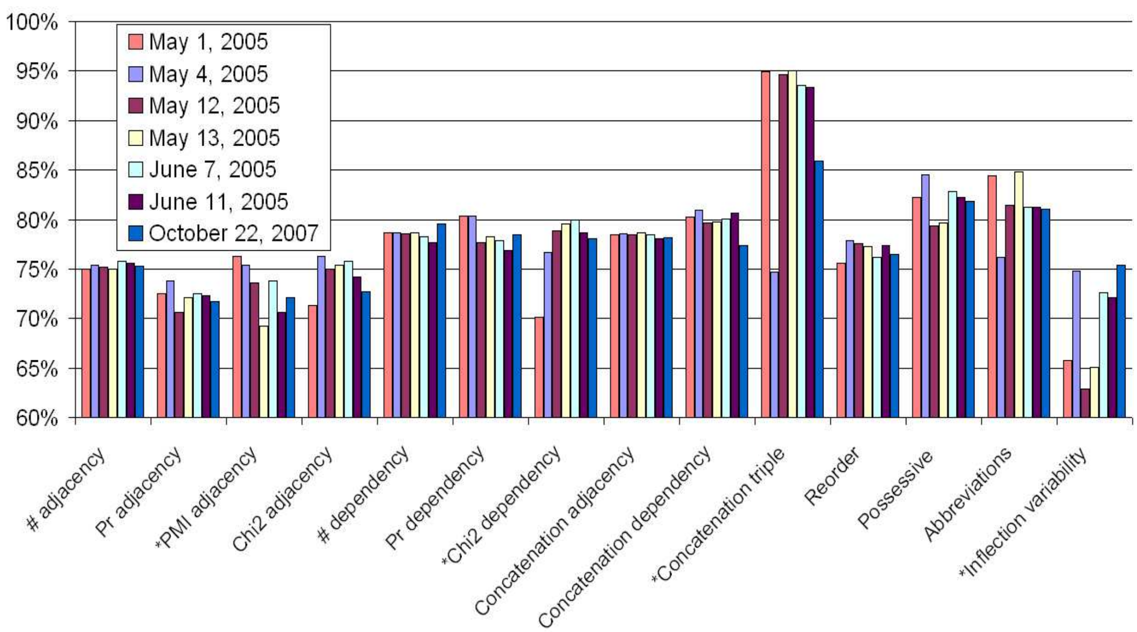

Figure 7.10: **Accuracy over time for *MSN*:** pages in any language, with word inflections. The models with statistically significant variations are marked with an asterisk.

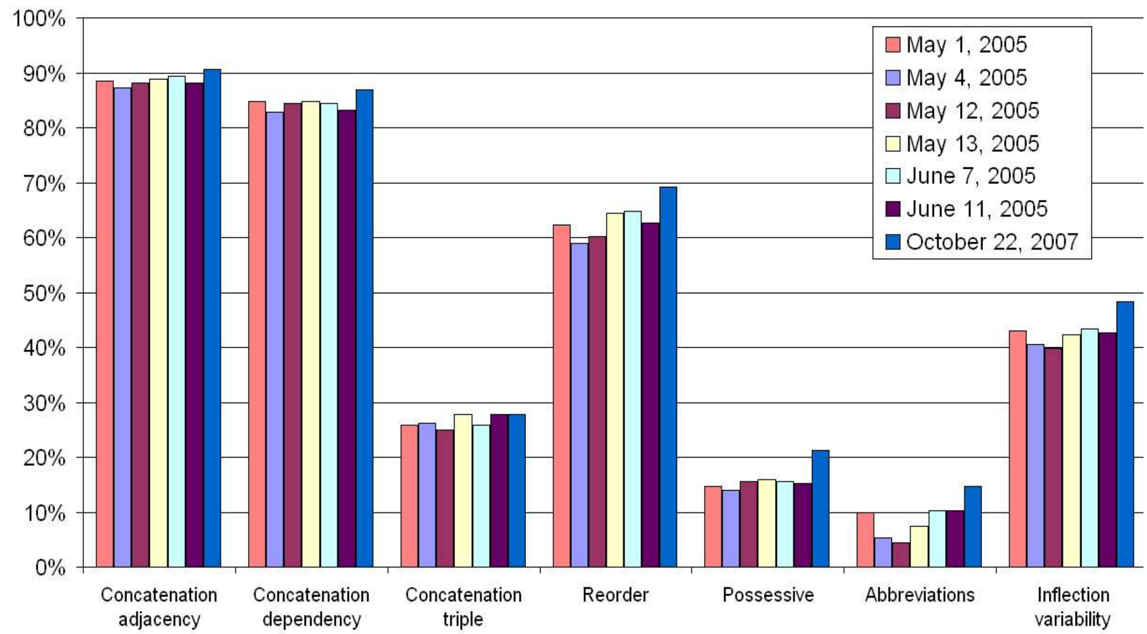

Figure 7.11: **Coverage over time for *MSN*:** pages in any language, no word inflections.

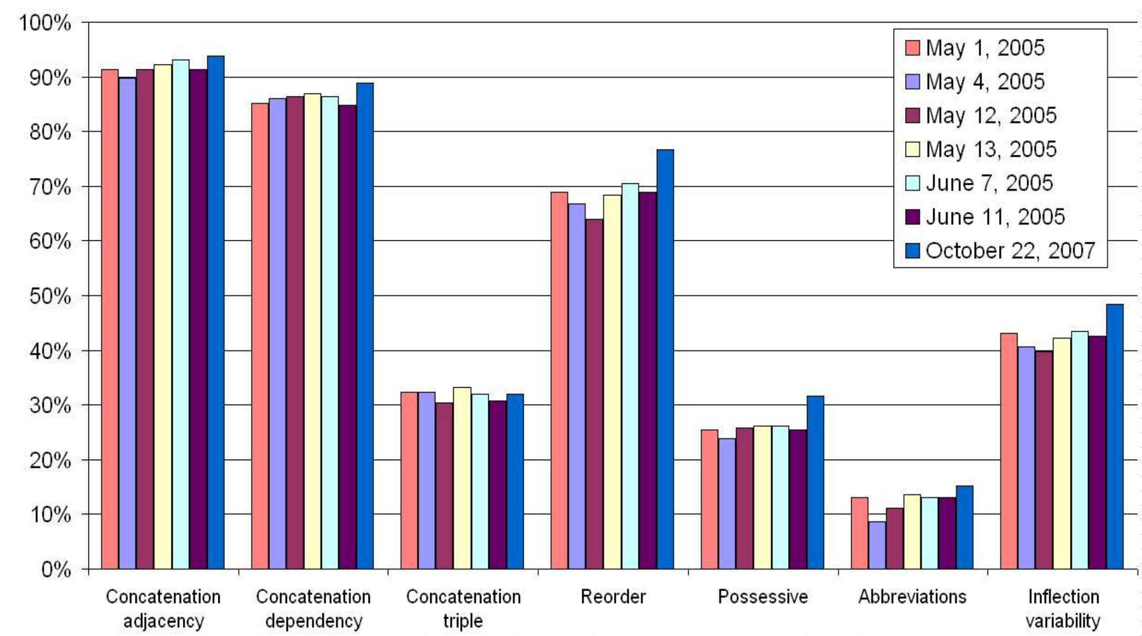

Figure 7.12: **Coverage over time for *MSN*:** pages in any language, with word inflections.

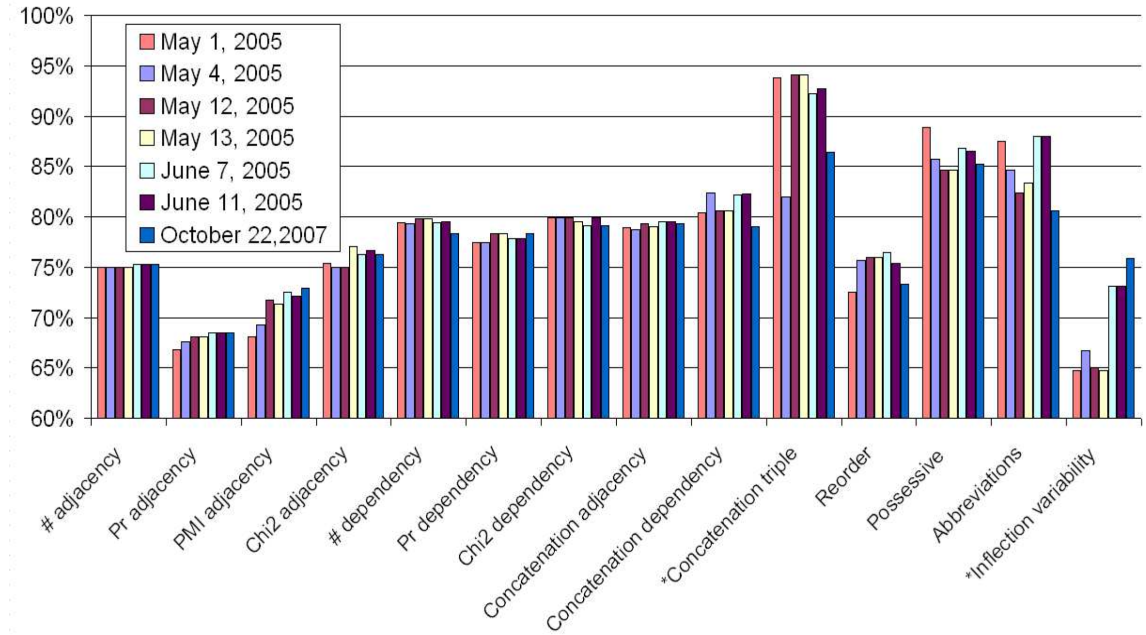

Figure 7.13: **Accuracy over time for *MSN*:** English pages only, no word inflections. The models with statistically significant variations are marked with an asterisk.

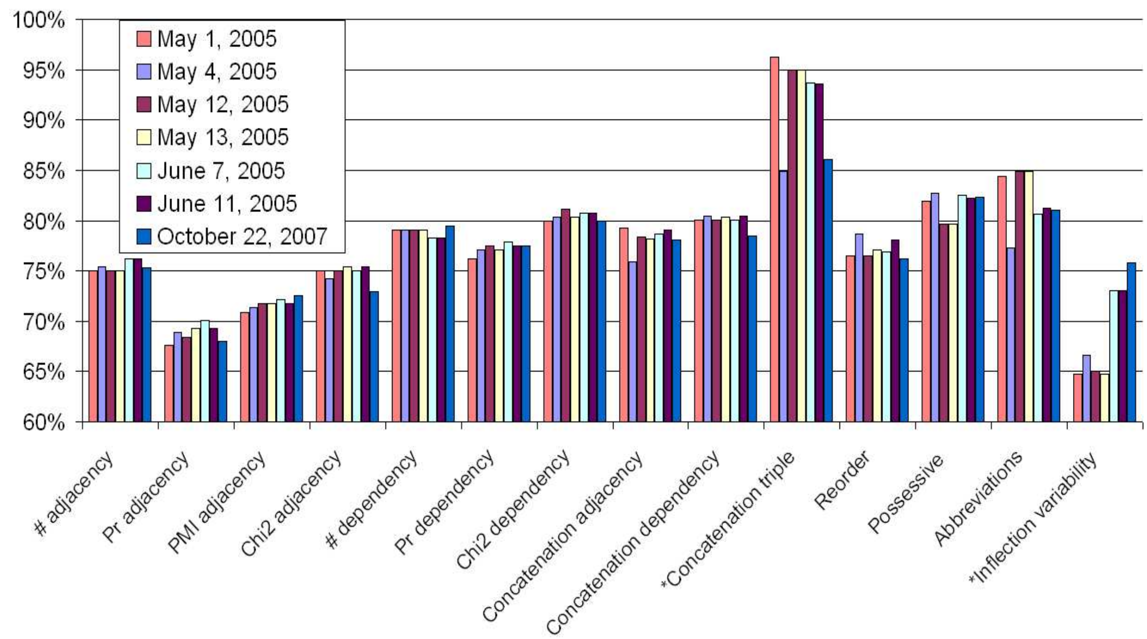

Figure 7.14: **Accuracy over time for *MSN*:** English pages only, with word inflections. The models with statistically significant variations are marked with an asterisk.

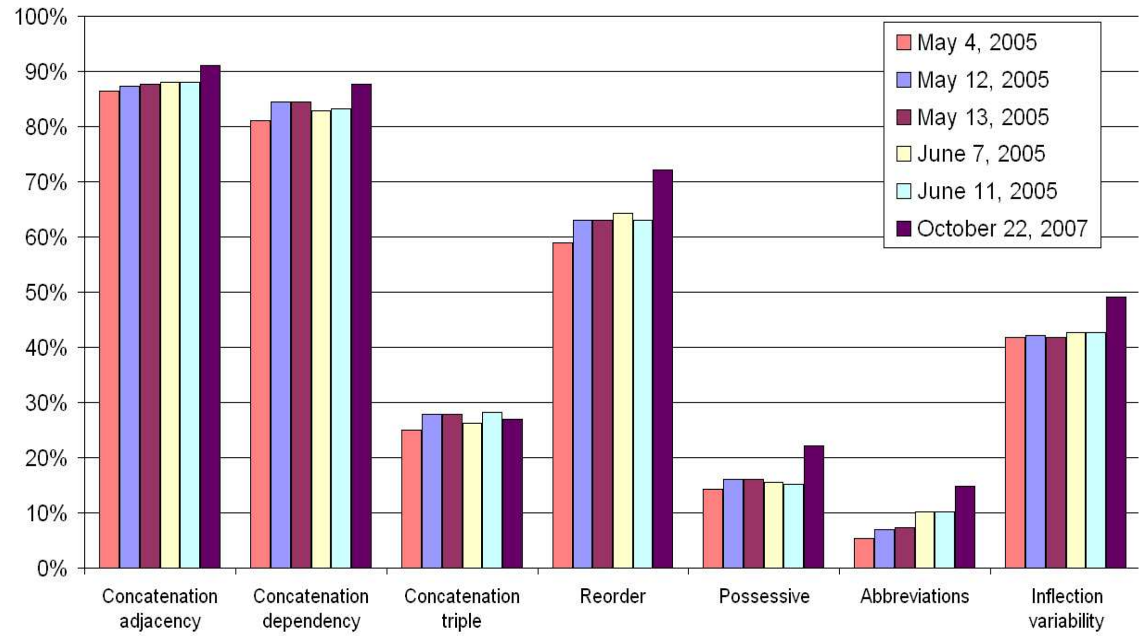

Figure 7.15: **Coverage over time for *MSN*:** English pages only, no word inflections.

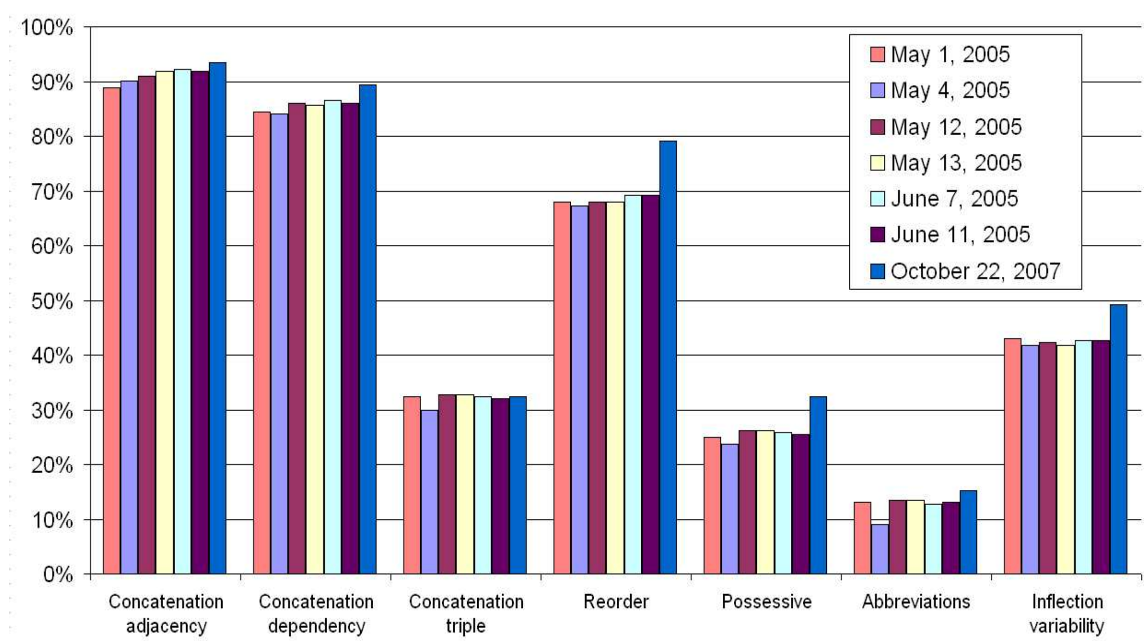

Figure 7.16: **Coverage over time for *MSN*:** English pages only, with word inflections.

nine models (about 1-4%), and more sizable for the last five ones. *Concatenation triple* is marked with an asterisk, which indicates statistically significant differences for it (see section 3.9.5). In Figure 7.6, there is a statistically significant difference for *PMI adjacency* and for *possessive* as well. While in the first two time snapshots the accuracy for the *possessive* is much lower than in the last two, it is the reverse for *concatenation triple*; *Google* may have changed how it handles possessive markers and word concatenations in the interrum.

The coverage (% of examples for which a model makes predictions) for the last seven models is shown in Figure 7.3 (no language filter, no word inflections), Figure 7.4 (no language filter, using word inflections), Figure 7.7 (English pages only, no word inflections) and Figure 7.8 (English pages only, using word inflections). The maximum difference in coverage for these models is rather small and does not exceed 7%. The coverage for the remaining seven models is not shown since it is 100% (or very close to 100%) for all models and under all observed conditions.

For *MSN Search*, I collected n-gram statistics for October 22, 2007 and for six dates in 2005: May 1, May 4, May 12, May 13, June 7, and June 11. The results for the accuracy are shown in Figure 7.9 (no language filter, no word inflections), Figure 7.10 (no language filter, using word inflections), Figure 7.13 (English pages only, no word inflections) and Figure 7.14 (English pages only, using word inflections). The corresponding results for the coverage are shown in Figures 7.11, 7.12, 7.15, 7.16. As these figures show, *MSN Search* exhibits much higher variability in accuracy compared to *Google*. There is almost 10% difference for the adjacency- and dependency-based models: for example, for χ^2 *dependency*, Figure 7.10 shows an accuracy of 70.08% for May 1, 2005 and 79.92% for June 7, 2005. This difference is statistically significant (and so is the difference for *PMI adjacency*: 76.23% on May 1, 2005 vs. 69.26% on May 13, 2005). Even bigger differences can be observed for the last seven models, e.g., Figure 7.9 shows about 14% absolute variation for *inflection variability* (75.42% on October 22, 2005 vs. 62.89% on May 12, 2007), and over 20% for the *concatenation triple* (74.68% on May 4, 2005 vs. 95.06% on May 13, 2005); both differences are statistically significant. The higher variability in accuracy for *MSN Search* is probably due to differences in how rounding is computed: since *Google*'s page hits are rounded, they change less over time. By contrast, these counts are exact for *MSN Search*, which makes them more sensitive to variation. This hypothesis is supported by the fact that overall both engines exhibit a higher variability for the last seven models, which use smaller counts that are more likely to be represented by exact numbers in *Google* as well.

Finally, the variability in coverage for *MSN Search* is slightly bigger than for *MSN Search* and goes up to 11% (as opposed to up to 7% for *Google*).

7.2.2 Variability by Search Engine

I study the variability by search engine by comparing *Google*, *MSN Search* and *Yahoo!* on the same day, June 6, 2005, in order to minimize the impact of index changes over time. I show the variability in accuracy by search engine and word inflection in Figure 7.17 (without language filters) and Figure 7.18 (English pages only). The corresponding results for the coverage are shown in Figures 7.19 and 7.20. I also show the variability in accuracy by search engine and language in Figure 7.21 (without word inflections) and Figure 7.22 (using word inflections); the coverage is shown in Figures 7.23 and 7.24.

As Figure 7.22 shows, the biggest difference in accuracy is exhibited by *concatena-*

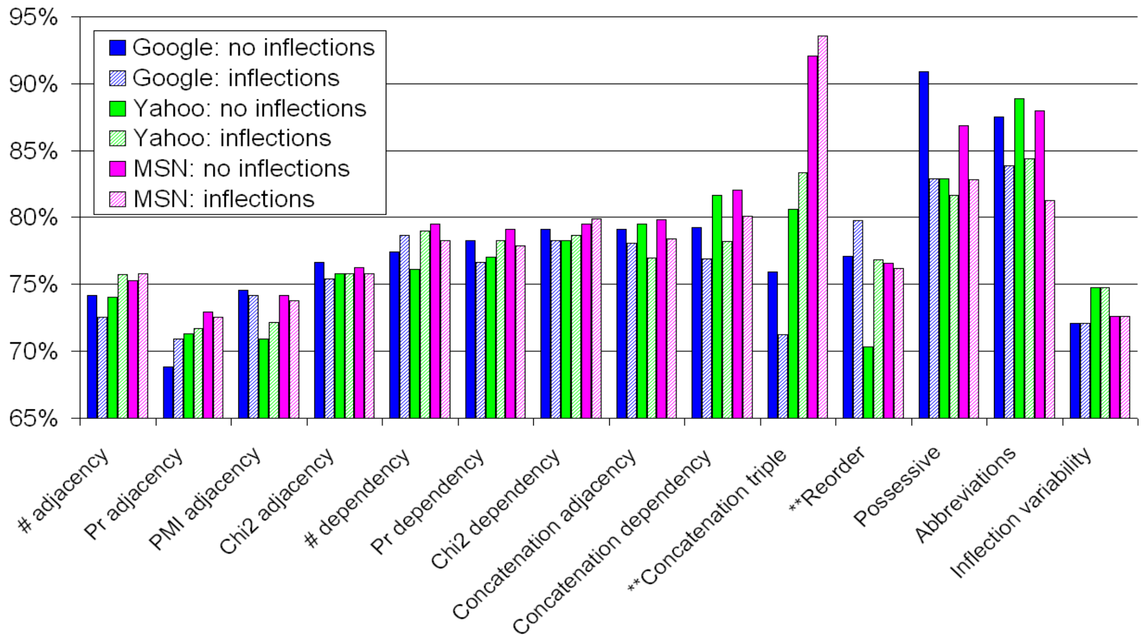

Figure 7.17: **Accuracy by search engine and inflection, June 6, 2005:** any language. Significant differences between different engines are marked with a double asterisk.

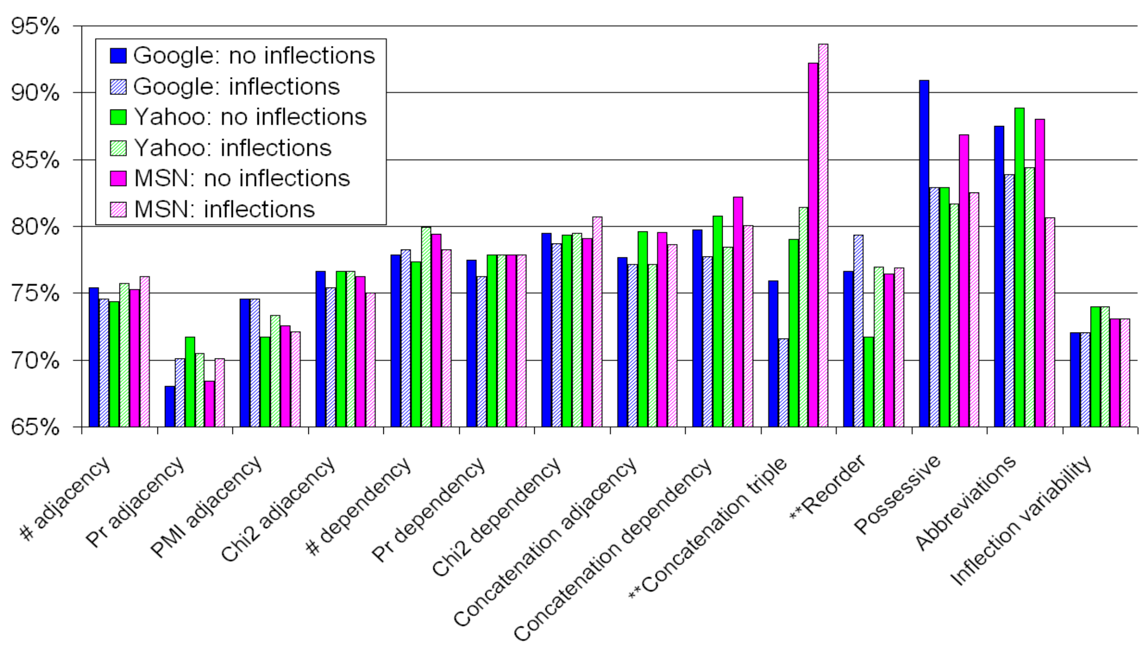

Figure 7.18: **Accuracy by search engine and inflection, June 6, 2005:** English only. Significant differences between different engines are marked with a double asterisk.

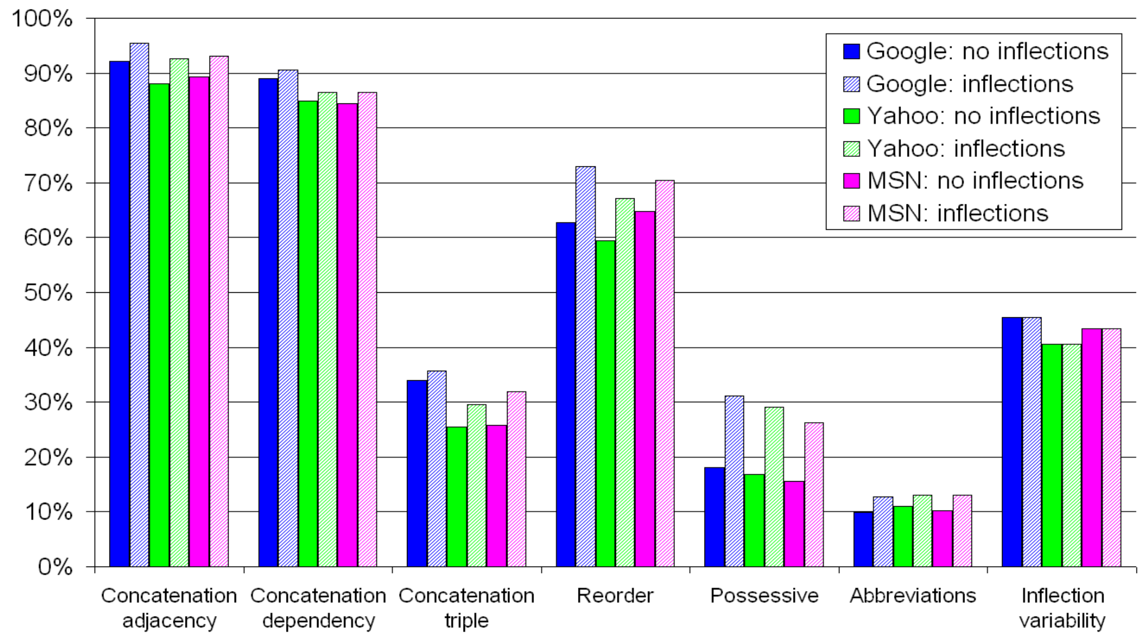

Figure 7.19: **Coverage by search engine and inflection, June 6, 2005:** any language.

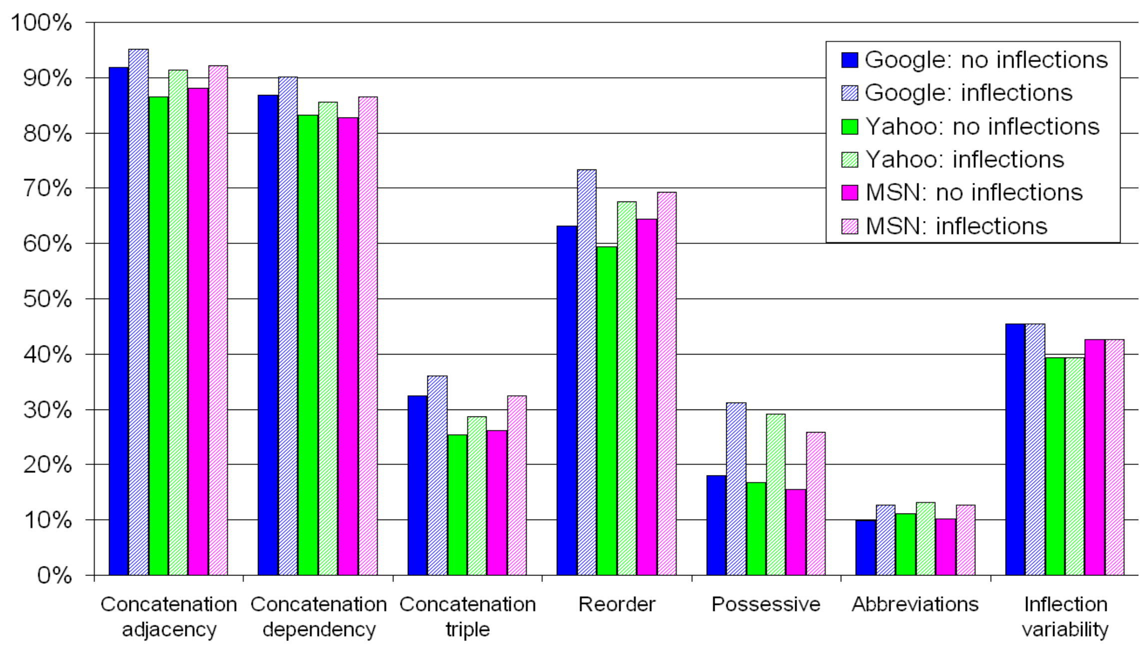

Figure 7.20: **Coverage by search engine and inflection, June 6, 2005:** English only.

tion triple: using inflected queries and English pages only, *MSN Search* achieves an accuracy of 93.67%, while *Google* under the same conditions only achieves 71.59%, which is a statistically significant difference. The results without the language filter are similar: 93.59% vs. 71.26%. The differences between *MSN Search* and *Google* are also statistically significant without the language filter. The differences for *reorder* between *Google* and *Yahoo!* are statistically significant as well. Other large variations (not statistically significant) can be seen for *possessive*. Overall, *MSN Search* performs best, which I attribute to it not rounding its page hit estimates. However, as Figure 7.21 shows, *Google* is almost 5% ahead on *possessive* for non-inflected queries, while *Yahoo!* leads on *abbreviations* and *inflection variability*. The fact that different search engines exhibit strength on different kinds of queries and models suggests that combining them could offer potential benefits. The results for the coverage in Figures 7.23 and 7.24 are slightly better for *Google*, which suggests it might have had a bigger index on June 6, 2005, compared to *Yahoo!* and *MSN Search*.

7.2.3 Impact of Language Filtering

The impact of language filtering, meaning requiring only documents in English as opposed to having no restrictions on the language, can be observed on several graphs.

First, the variation in accuracy can be observed across search engines for June 6, 2005, as shown in Figure 7.21 (without word inflections) and Figure 7.22 (using word inflections). The impact of language filtering on accuracy is minor and inconsistent for all three search engines: sometimes the results are improved slightly and sometimes they are negatively impacted, with the differences varying up to 3% (usually 0-1%). Figures 7.23 and 7.24, compare the impact of language filtering on coverage for the last seven models.

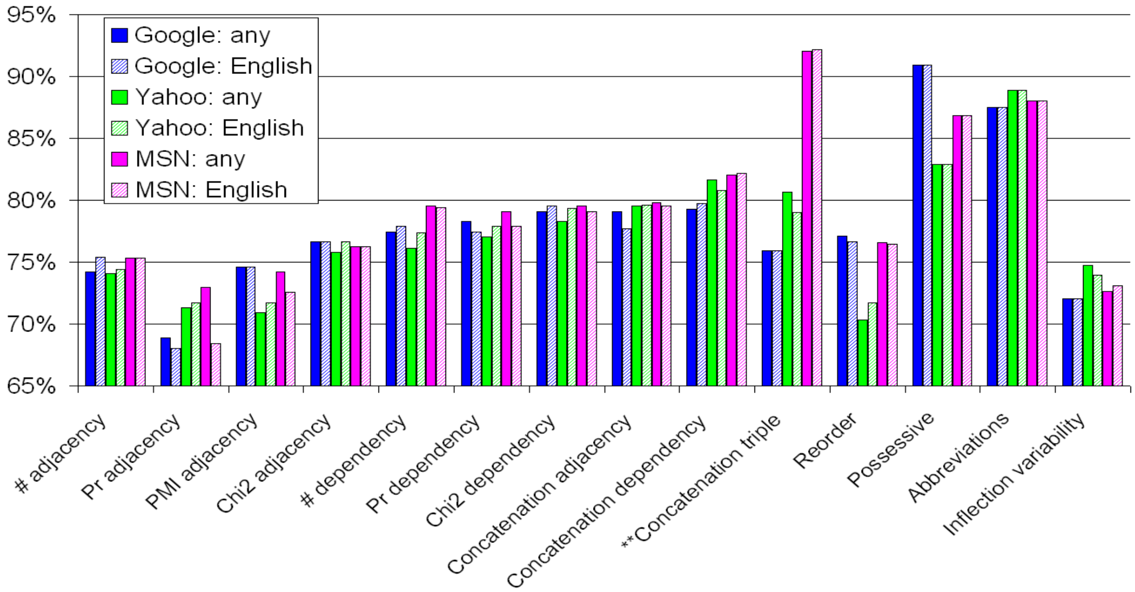

Figure 7.21: **Accuracy by search engine and language filtering, June 6, 2005:** no inflections. Significant differences between different engines are marked with a double asterisk.

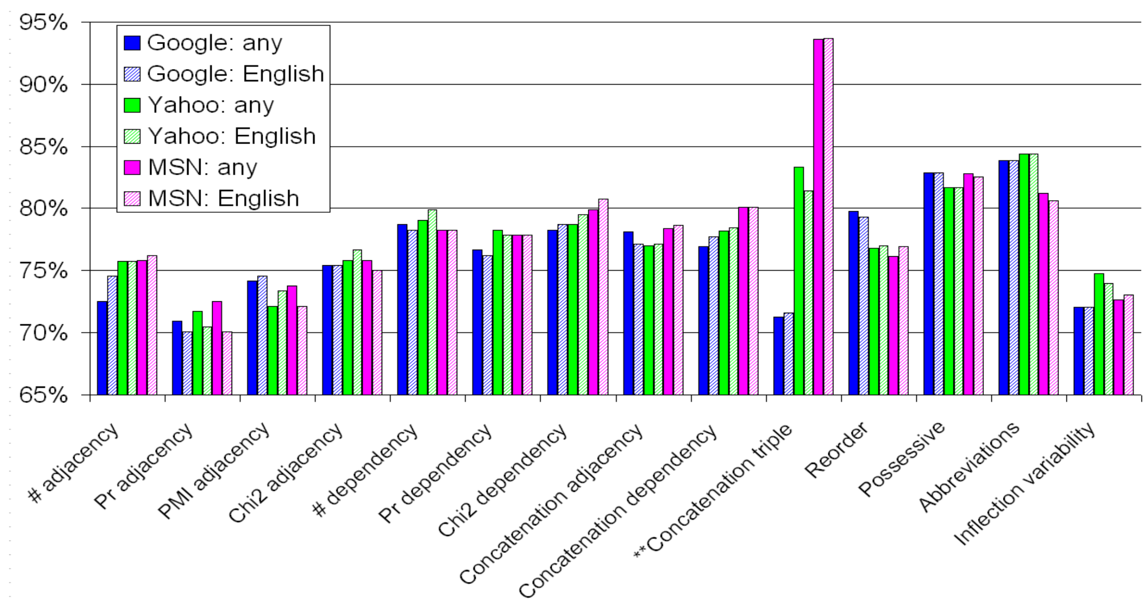

Figure 7.22: **Accuracy by search engine and language filtering, June 6, 2005:** with inflections. Significant differences between different engines are marked with a double asterisk.

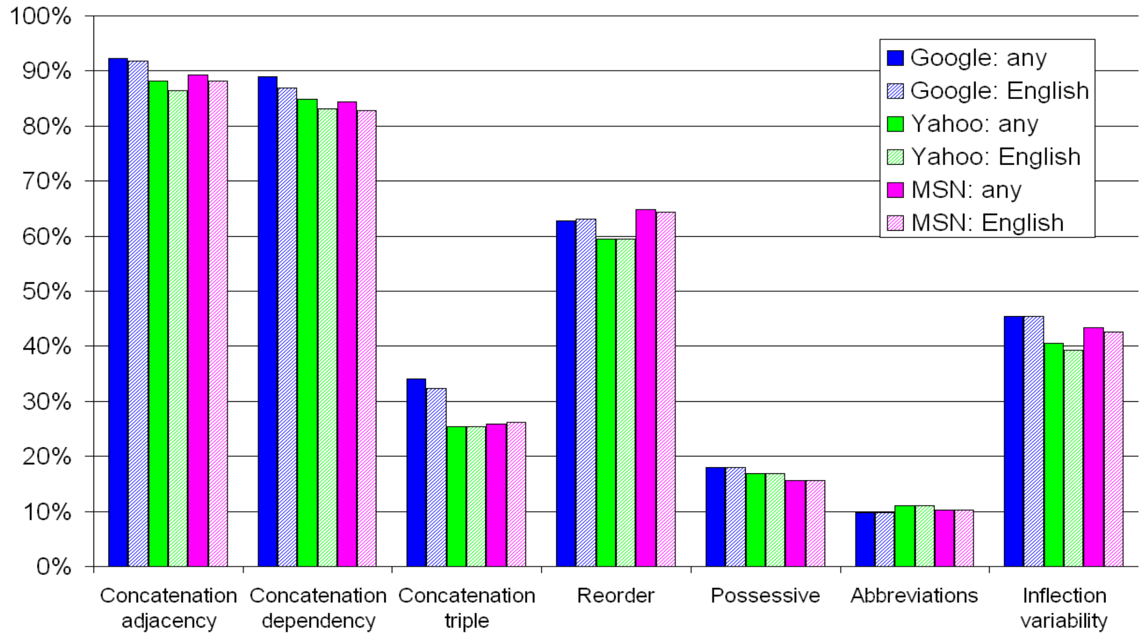

Figure 7.23: **Coverage by search engine and language filtering, June 6, 2005:** no inflections.

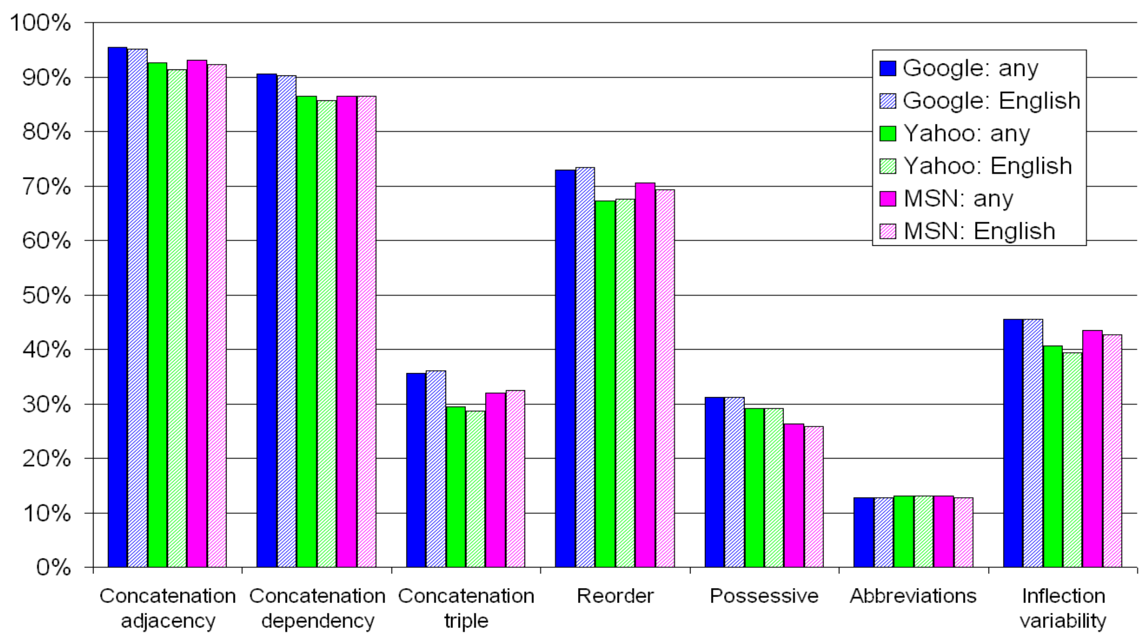

Figure 7.24: **Coverage by search engine and language filtering, June 6, 2005:** with inflections.

As we can see, using English only leads to a drop in coverage, as one could expect, but this drop is quite small: less than 1%.

Second, the impact of language filtering can also be observed across time for *Google* and for *MSN Search*. For *Google*, we can study the impact without inflections by comparing Figure 7.1 (no language filter) and Figure 7.5 (English pages only). We can also study the impact with inflections by comparing Figure 7.2 (no language filter) and Figure 7.6 (English pages only). Similarly, for *MSN Search*, the impact without inflections would compare Figure 7.9 (no language filter) and Figure 7.13 (English pages only). We can also study the impact with inflections by comparing Figure 7.10 (no language filter) and Figure 7.14 (English pages only). In all cases, the differences in accuracy are very small: less than 3%. The corresponding differences in coverage are within 0-1%.

7.2.4 Impact of Word Inflections

Finally, I study the impact of using different word inflections and adding up the frequencies. I generate the inflections (e.g., *tumor* and *tumors*) using *Carroll's morphological tools* (Minnen *et al.* 2001) and *WordNet* (Fellbaum 1998), as described in section 3.9. See sections 3.4 and 3.5 for a detailed description of what kinds of inflections are used for each model and how the page hits are summed up.

The impact of word inflections can be studied by looking at several graphs. First, the variation in accuracy can be observed across search engines for June 6, 2005, as shown in Figure 7.17 (no language filter) and Figure 7.18 (English only). While the results are mixed, the impact on accuracy here is bigger compared to language filtering (and, of course, there is no impact on *inflection variability*). Interestingly, for all search engines and regardless of whether language filtering has been used, inflected queries positively impact the accuracy for *possessive* (by up to 8%), *abbreviations* (by up to 8%), *concatenation adjacency* (by up to 3-4%) and *concatenation dependency* (by up to 4%). Using word inflections for *concatenation triple* improves accuracy by 2-4% for *Yahoo!* and *MSN Search*, but negatively affects *Google* (by about 5%). Overall, word inflections improve accuracy for *reorder* by about 7% for *Yahoo!*, and by about 3% for *Google*, but have an insignificant impact for *MSN Search*. The impact of using word inflections on coverage can be seen in Figure 7.19 (no language filter) and Figure 7.20 (English only). As one would expect, the coverage improves consistently (by up to 7%), especially for *possessive*, *reorder* and *concatenation triple*. However, none of these differences in accuracy or coverage is statistically significant.[3]

Second, the impact of using word inflections can be observed across time for *Google* and for *MSN Search*. For *Google*, the impact of word inflections, in case of no language filtering, can be seen by comparing Figure 7.1 (without word inflections) and Figure 7.2 (using word inflections), while in case of filtering out the non-English pages we can compare Figure 7.5 (without word inflections) and Figure 7.6 (using word inflections). Similarly, for *MSN Search*, the impact of word inflections, without language filtering, can be seen by comparing Figure 7.9 (without word inflections) and Figure 7.10 (using word inflections), while in case of filtering out the non-English pages we can compare Figure 7.13 (without word inflections) and Figure 7.14 (using word inflections). In either case, the differences in accuracy are up to 8%, and the corresponding differences in coverage are up to 7%; none of these variations is statistically significant.

[3]Figures 7.17 and 7.18 only contain double asterisks, which indicate statistically significant difference between search engines, not between inflected and non-inflected queries.

7.2.5 Variability by Search Engine: 28 Months Later

I have studied the interesting question of how the variability by search engine changes over a long period of time for different search engines. For the purpose, I repeated the experiments described in section 7.2.2 comparing the performance of *Google*, *Yahoo!* and *MSN* on June 6, 2005 and on October 22, 2007.

Many things have changed over this period of time: the size of search engines' indexes has increased query analysis has been updated (e.g., *Google* seems to handle automatically inflectional variation for English, which has an impact on inflected queries), and page hit estimation formulas employed by search engines have been altered. The most important change occurred to *MSN Search*, now *Live Search*, which started rounding its page hit estimates, just like *Google* and *Yahoo!* do, rather than providing exact estimates as in 2005.

The evaluation results are shown on the following figures: Figure 7.25 (no language filter, no word inflections), Figure 7.26 (no language filter, using word inflections), Figure 7.29 (English pages only, no word inflections) and Figure 7.30 (English pages only, using word inflections). We can see that, while the differences in performance are inconsistent, the accuracies for *Live Search* on October 22, 2007 are generally worse compared to those for *MSN Search* on June 6, 2005, especially for the last six models, which I attribute to *Live Search* rounding its page hit estimates.

The corresponding results for the coverage are shown in Figures 7.27, 7.28, 7.31, 7.32. As we can see, the coverage has improved over time for all three search engines, which can be expected, given the increase in the number of indexed Web pages.

7.2.6 Page Hit Estimations: Top 10 vs. Top 1,000

As I mentioned in section 7.1.4, search engines may sample from their indexes, rather than performing exact computations, in which case index scanning might stop once the requested number of pages is produced: 10 by default, and up to 1,000 at most. The more pages the user requests, the larger the part of the indexes scanned, and the more accurate the estimates would be.

To see the impact of the more accurate estimates, I extract the page hit estimates after having asked the search engine to produce 10 and 1,000 results. The corresponding accuracies for *Google* and *Yahoo!* on October 22, 2007 are shown in Figure 7.33 (no inflections, any language) and Figure 7.34 (with inflections, any language). Interestingly, using estimates after having requested 1,000 results has a negative impact on accuracy, i.e. the initial page hit estimates were actually better. Note also that *Yahoo!* shows less variability in accuracy than *Google*, which suggests *Yahoo!*'s page hit estimates are more stable.

7.3 Conclusion and Future Work

In the present chapter, I have described some problems with using search engine page hits as a proxy for n-gram frequency estimates. Then, using a real NLP task, *noun compound bracketing*, and 14 different n-gram based models, I have shown that variability over time and across search engines, as well as using language filters and word inflections, while sometimes causing sizable fluctuations, generally have no statistically significant impact on accuracy and coverage for that particular task (except for "exotic" models like *concatenation triple*, *inflection variability*, etc. under some of the experimental conditions).

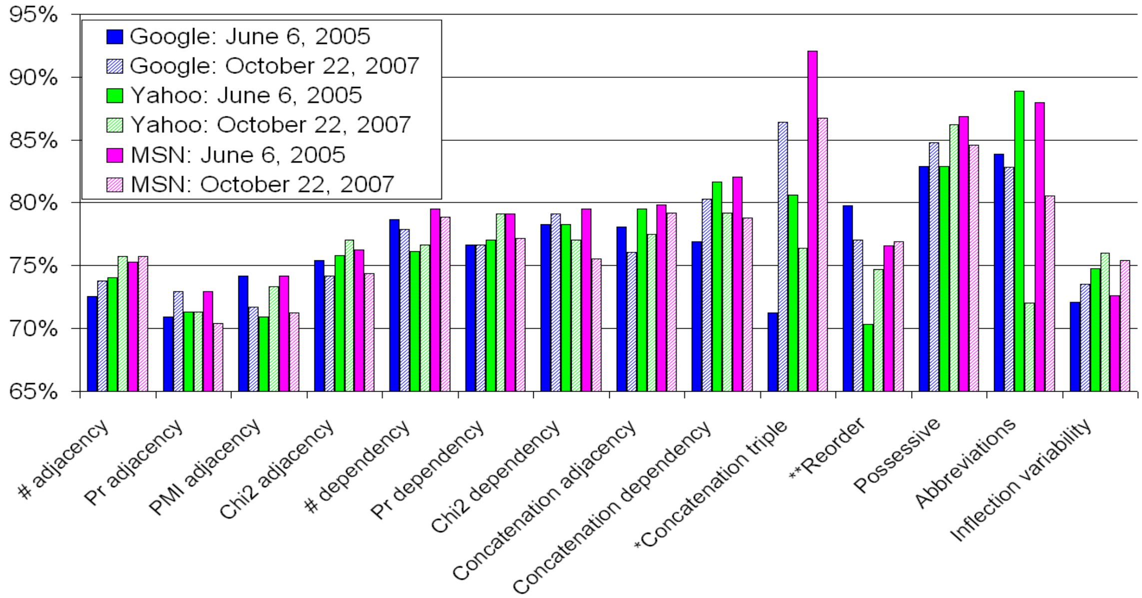

Figure 7.25: **Accuracy by search engine for June 6, 2005 and October 22, 2007:** no inflections, any language. Statistically significant differences between the same/different search engines is marked with a single/double asterisk.

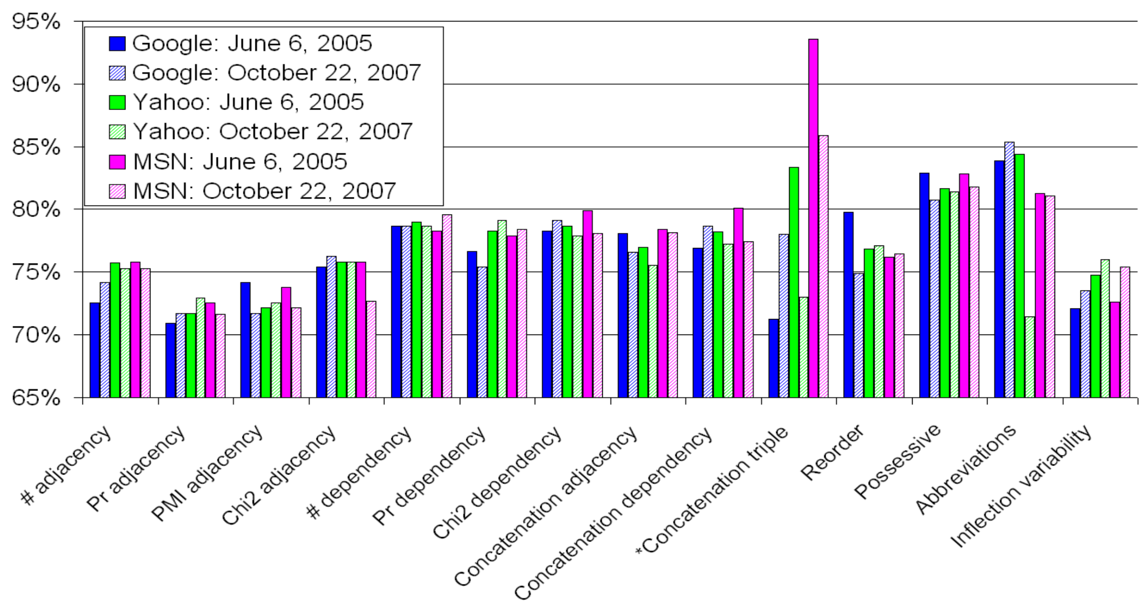

Figure 7.26: **Accuracy by search engine for June 6, 2005 and October 22, 2007:** using inflections, any language. Statistically significant differences between the same search engine are marked with an asterisk.

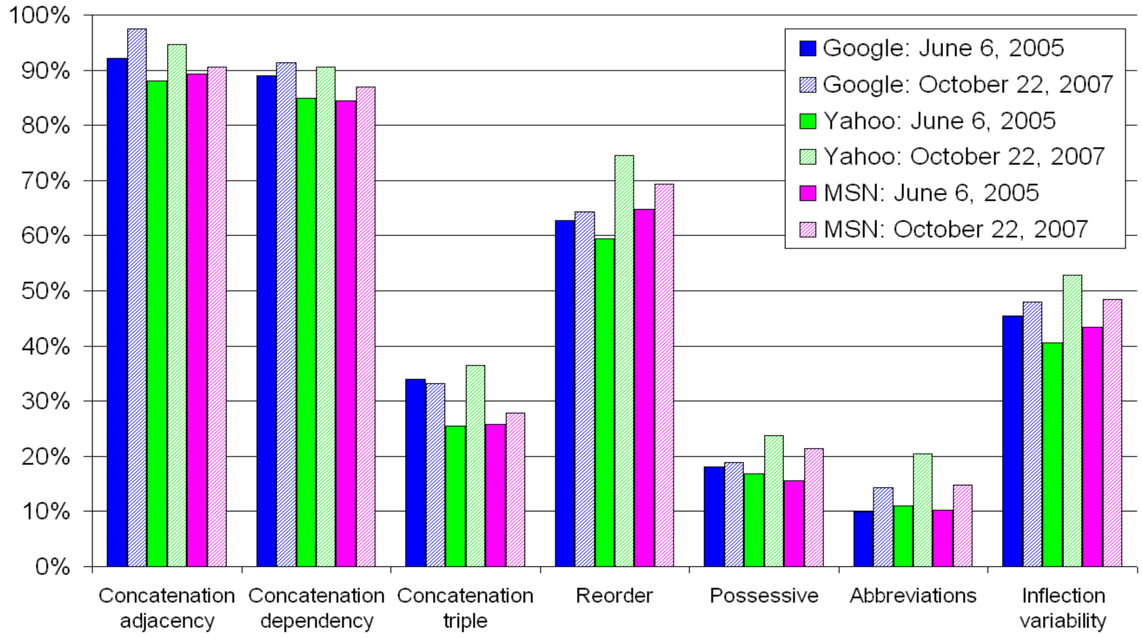

Figure 7.27: **Coverage by search engine for June 6, 2005 and October 22, 2007:** no inflections, any language.

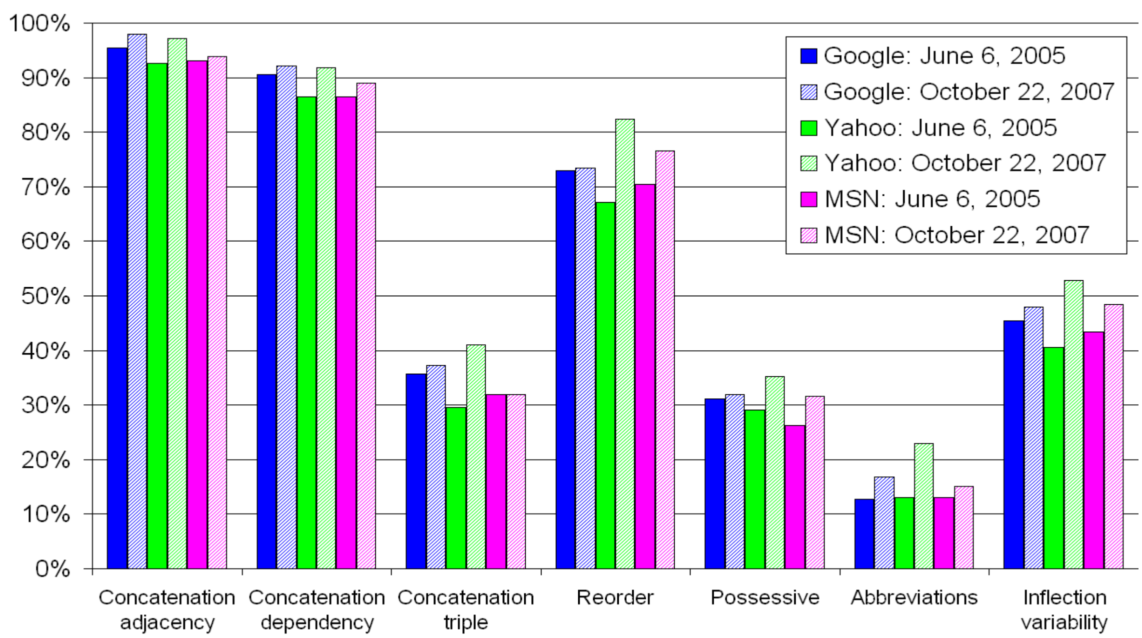

Figure 7.28: **Coverage by search engine for June 6, 2005 and October 22, 2007:** using inflections, any language.

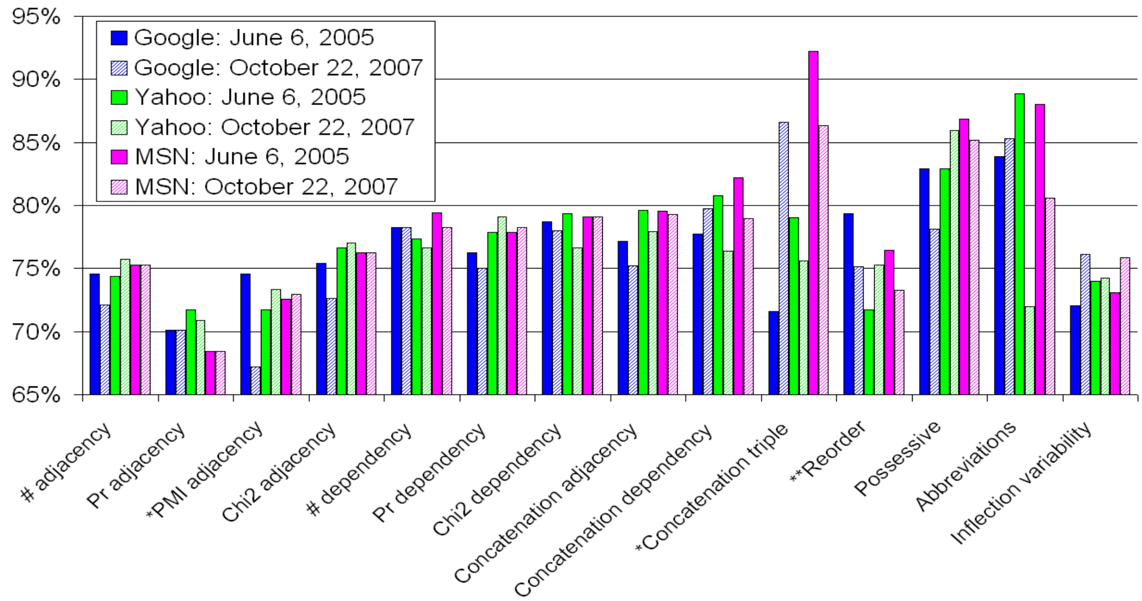

Figure 7.29: **Accuracy by search engine for June 6, 2005 and October 22, 2007:** no inflections, English pages only. Statistically significant differences between the same/different search engines is marked with a single/double asterisk.

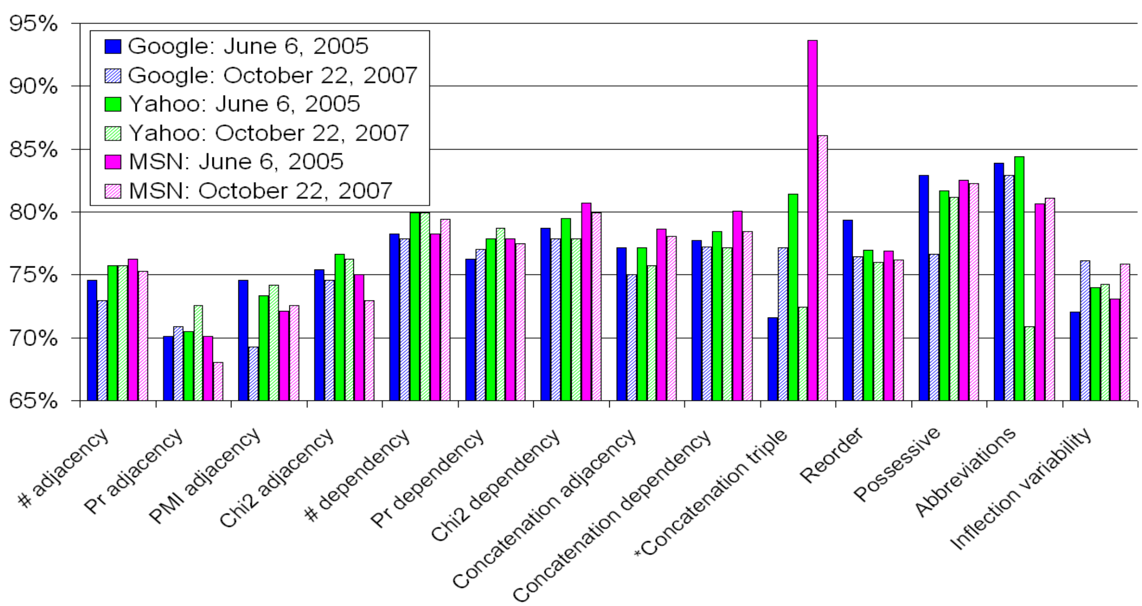

Figure 7.30: **Accuracy by search engine for June 6, 2005 and October 22, 2007:** using inflections, English pages only. Statistically significant differences between the same search engine are marked with an asterisk.

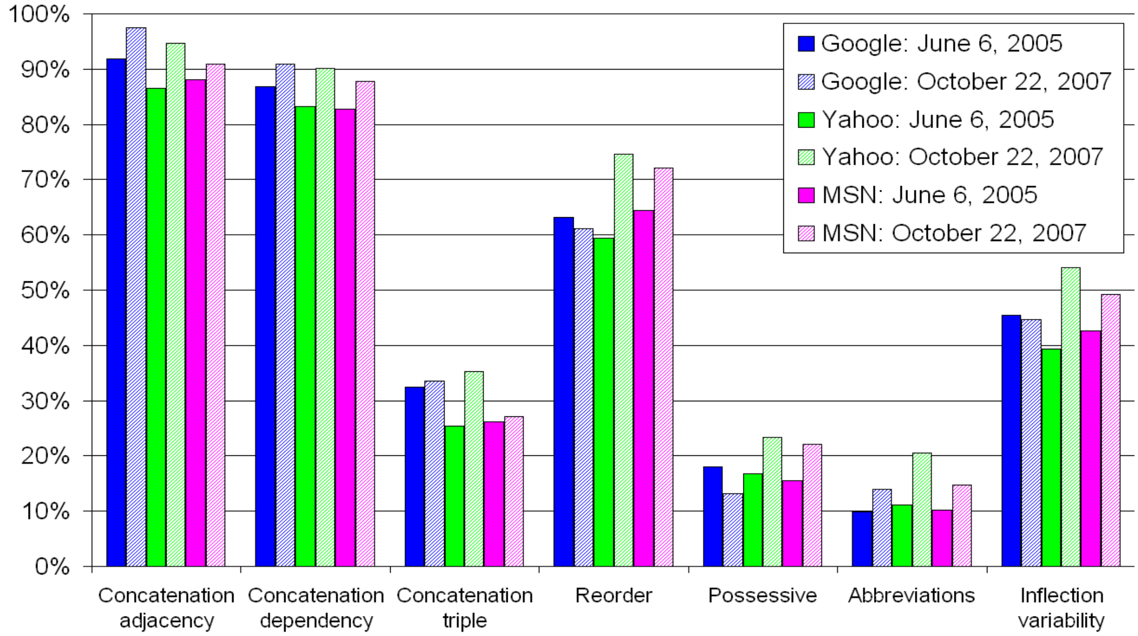

Figure 7.31: **Accuracy by search engine for June 6, 2005 and October 22, 2007:** no inflections, English pages only.

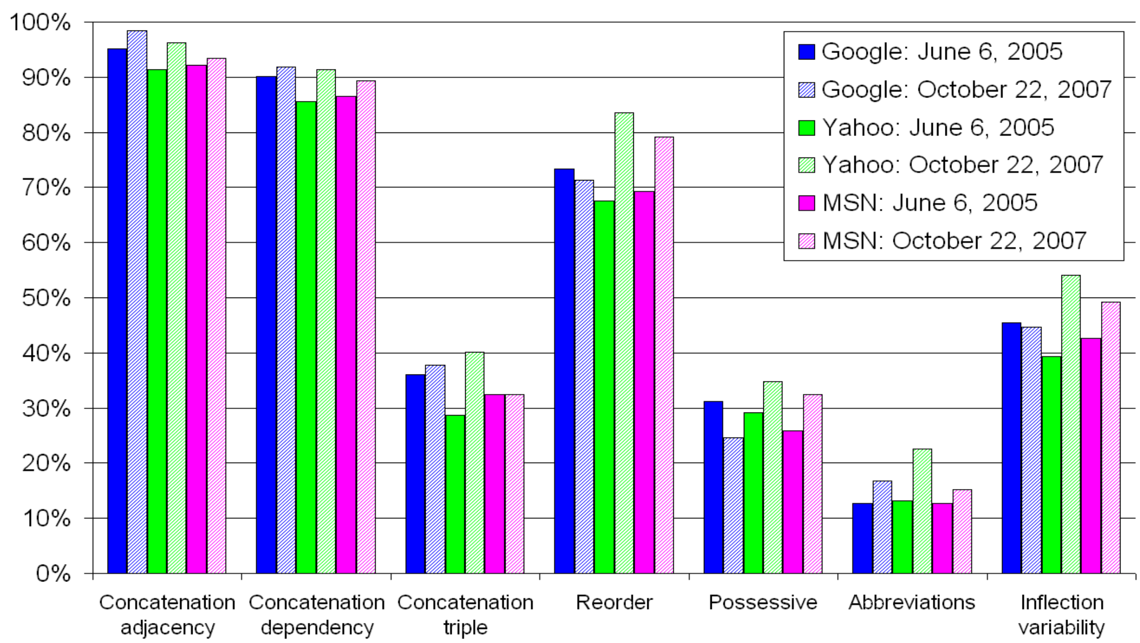

Figure 7.32: **Accuracy by search engine for June 6, 2005 and October 22, 2007:** using inflections, English pages only.

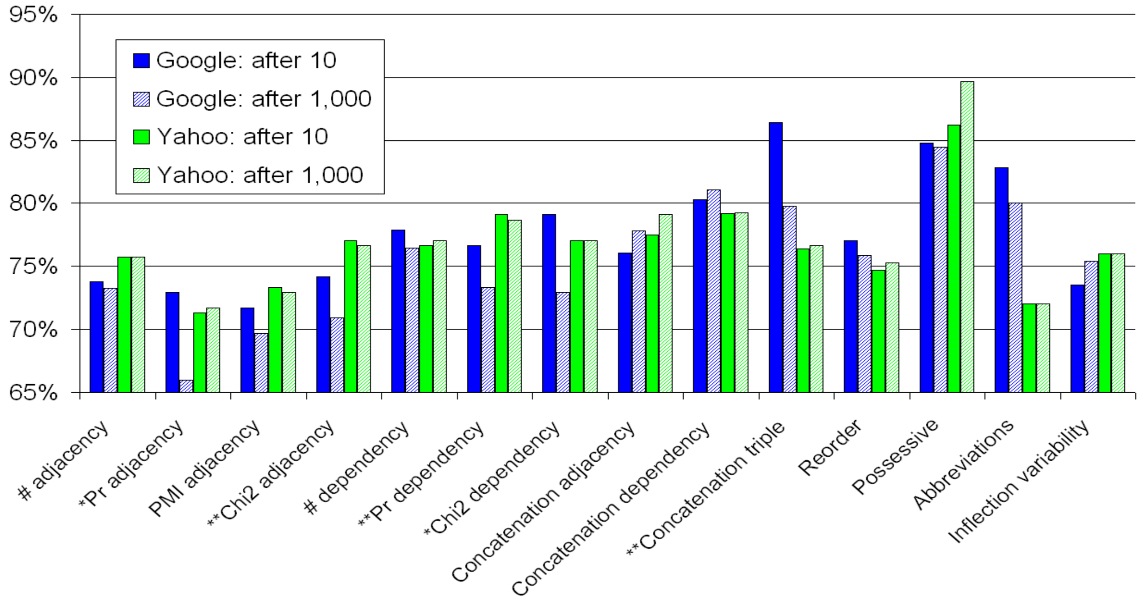

Figure 7.33: **Accuracy by search engine for October 22, 2007, 10 vs. 1,000 results:** no inflections, any language. Statistically significant differences between the same/different search engines is marked with a single/double asterisk.

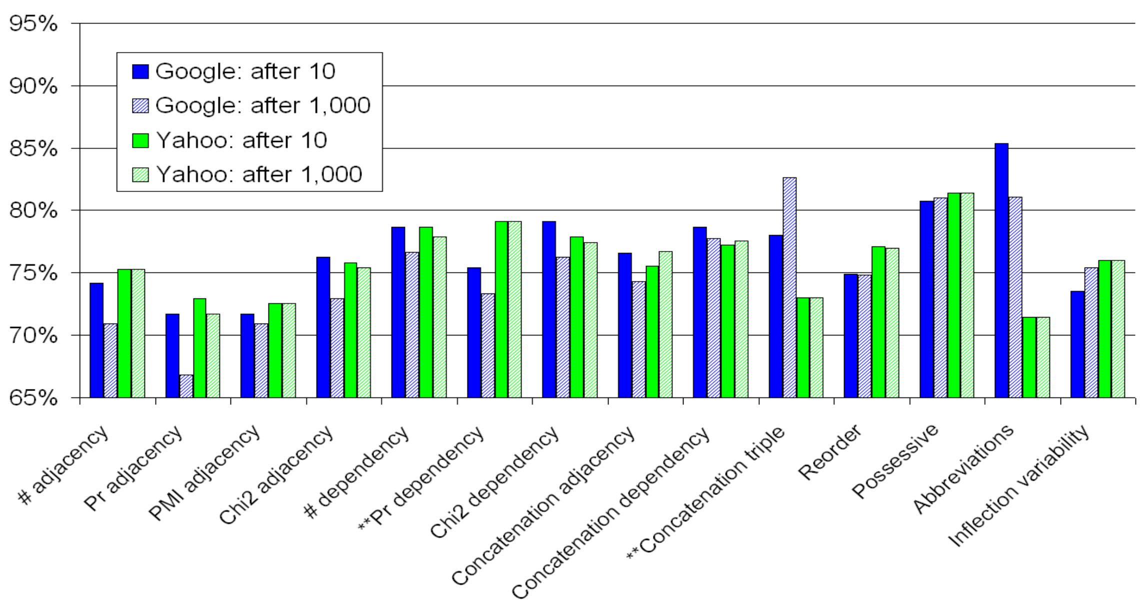

Figure 7.34: **Accuracy by search engine for October 22, 2007, 10 vs. 1,000 results:** with inflections, any language. Statistically significant differences between the same/different search engines is marked with a single/double asterisk.

These results suggest that, despite the fluctuations, overall experiments with models using search engines statistics can be considered relatively stable and reproducible. This is good news which should reassure researchers using Web as a data source that, despite the recent scepticism of Kilgarriff (2007), *Googleology* is not "bad science" after all.

In order to further bolster these results, more experiments with other NLP tasks which make use of Web-derived n-gram estimates would be needed. It would be good to perform a similar study for *surface features* (see chapter 3.5) and for *paraphrases* (see chapter 3.7) as well. It would be also interesting to try languages other than English, where the language filter could be much more important, and where the impact of the inflection variability may differ, especially in case of a morphologically rich language like Bulgarian or Russian. Finally, since different experimental conditions, e.g., different search engines, exhibit strength for different kinds of models, it looks promising to combine them, e.g., in a log-linear model, similarly to the way Lapata & Keller (2005) combined n-gram statistics from the Web and from the *British National Corpus*.

Chapter 8

Conclusions and Future Work

8.1 Contributions of the Thesis

In the present thesis, I have introduced novel surface features and paraphrases which go beyond simple n-gram counts, which have been demonstrated highly effective for a wide variety of NLP tasks. Based on them, I have built several novel, highly accurate Web-based approaches to solving problems related to the syntax and semantics of noun compounds, and to other related problems like relational similarity, machine translation, PP-attachment and NP coordination.

In chapter 3, I have extended and improved upon the state-of-the-art approaches to noun compound bracketing, achieving 89.34% accuracy on the benchmark *Lauer's dataset*, which is a statistically significant improvement over the best previously published results.

In chapter 4, I have presented a novel, simple unsupervised approach to noun compound interpretation in terms of predicates characterizing the hidden relation between the head and the modifier, which could be useful for many NLP tasks. Using these verbs as features in a classifier, I have demonstrated state-of-the-art results for several relational similarity problems, including: mapping noun-modifier pairs to abstract relations like `TIME` and `LOCATION`, classifying relation between nominals, and solving SAT verbal analogy questions.

In chapter 5, I have described an application of the methods developed in chapters 3 and 4 for noun compound paraphrasing to an important real-world task: *machine translation*. I have proposed a novel monolingual paraphrasing method based on syntactic transformations at the NP-level, used to increase the training data with nearly equivalent sentence-level syntactic paraphrases of the original corpus. The evaluation has shown an improvement equivalent to 33%-50% of that of doubling the amount of training data.

In chapter 7, I have addressed the important question of the stability and reproducibility of the results obtained using search engines. Using a real NLP task, *noun compound bracketing*, I have shown that variability over time and across search engines, as well as using language filters and word inflections, generally has no statistically significant impact. These results suggest that, despite the fluctuations, overall experiments with models using search engines statistics can be considered stable and reproducible.

Finally, I have created three datasets which might be useful to other researchers:

(1) a domain-specific *Biomedical dataset* for noun compound bracketing is given in Appendix C.3;

(2) an *NP-coordination dataset* is listed in Appendix D;

(3) a distribution of the top 10 human-generated verbs paraphrasing the hidden relation between the nouns in 250 noun-noun compounds used in the theory of Levi (1978) is given in Appendix E.

8.2 Directions for Future Research

In my opinion, the most exciting and probably the most promising direction for future research is on using the generated verbs and paraphrases of noun-noun compounds for various related NLP tasks, e.g., paraphrase-augmented machine translation (Callison-Burch *et al.* 2006), noun compound translation (Baldwin & Tanaka 2004), machine translation evaluation (Russo-Lassner *et al.* 2005; Kauchak & Barzilay 2006), summarization evaluation (Zhou *et al.* 2006), textual entailment (see Appendix A), information retrieval, e.g. index normalization (Zhai 1997), query segmentation (Bergsma & Wang 2007). relational Web search (Cafarella *et al.* 2006), lexical acquisition, e.g., extracting names of diseases and drugs from the Web (Etzioni *et al.* 2005), etc.

My future plans for the noun compound bracketing problem include extending the approach to noun compounds consisting of more than three nouns, bracketing NPs in general, and recognizing structurally ambiguous noun compounds. As the evaluation results in chapter 3 show, there may be potential benefits in combining Web and corpus-based statistics, e.g., similarly to the way Lapata & Keller (2005) combined n-gram estimates from the Web and from the *British National Corpus*. Another information source that seems promising for the syntax and semantics of noun compounds is cross-linguistic evidence, e.g., extracted from parallel multi-lingual corpora like *Europarl* as described by Girju (2007b).

It would be interesting to extend the page hit instability study from chapter 7, and to apply it to *surface features* (see section 3.5) and *paraphrases* (see section 3.7). Trying other NLP tasks would be interesting as well, as will be experimenting with languages other than English, where the language filter might be much more important, and where the impact of the inflection variability may be different, especially in case of a morphologically rich language like Bulgarian and Russian.

Last but not least, special efforts should be made to reduce the number of queries to the search engines. *Google*'s *Web 1T 5-gram* dataset, which is available from the Linguistic Data Consortium, while somewhat limited, might still be quite useful in this respect.

Bibliography

ABE, NAOKI, & HANG LI. 1996. Learning word association norms using tree cut pair models. In *In Proceedings of the International Conference on Machine Learning*, 3–11.

ABNEY, S., R.E. SCHAPIRE, & Y. SINGER. 1999. Boosting applied to tagging and PP-attachment. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP-VLC*, 38–45.

AGARWAL, RAJEEV, & LOIS BOGGESS. 1992. A simple but useful approach to conjunct identification. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, 15–21, Morristown, NJ, USA. Association for Computational Linguistics.

AGRESTI, A., & B. COULL. 1998. Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* 119–126.

AL-ONAIZAN, YASER, & KEVIN KNIGHT. 2001. Translating named entities using monolingual and bilingual resources. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 400–408, Morristown, NJ, USA. Association for Computational Linguistics.

ALEGRE, MARTHA A., JOSEP M. SOPENA, & AGUSTI LLOBERAS. 1999. PP-attachment: A committee machine approach. In *In Proceedings of EMNLP/VLC-99*, 231–238.

ALSHAWI, HIYAN, & DAVID CARTER. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics* 20.635–648.

ATANASSOVA, IVANKA. 2001. *Korrektorovychny kurs russkogo yazyka. (Grammaticheskiy aspekt). Posobie-spravochnik. pervaya chast' (in Russian).* Veliko Turnovo, Bulgaria: Faber.

BAKER, COLLIN F., CHARLES J. FILLMORE, & JOHN B. LOWE. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, 86–90, Morristown, NJ, USA. Association for Computational Linguistics.

——, & JOSEF RUPPENHOFER. 2002. FrameNet's frames vs. Levin's verb classes. In *In Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*, 27–38.

BALDWIN, TIMOTHY, & TAKAAKI TANAKA. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, 24–31.

BANKO, MICHELE, & ERIC BRILL. 2001. Scaling to very very large corpora for natural language disambiguation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 26–33, Morristown, NJ, USA. Association for Computational Linguistics.

BANNARD, COLIN, & CHRIS CALLISON-BURCH. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

BARKER, KEN, & STAN SZPAKOWICZ. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th international conference on Computational linguistics*, 96–102, Morristown, NJ, USA. Association for Computational Linguistics.

BARZILAY, REGINA, & KATHLEEN R. MCKEOWN. 2001. Extracting paraphrases from a parallel corpus. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 50–57, Morristown, NJ, USA. Association for Computational Linguistics.

BERGSMA, SHANE, & QIN IRIS WANG. 2007. Learning noun phrase query segmentation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 819–826.

BOYADZHIEV, TODOR, IVAN KUCAROV, & YORDAN PENCHEV. 1998. *Syvremenen bylgarski ezik (in Bulgarian)*. Sofia, Bulgaria: Izdatelska kyshta "Petar Beron".

BRILL, ERIC, & PHILIP RESNIK. 1994. A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th conference on Computational linguistics*, 1198–1204, Morristown, NJ, USA. Association for Computational Linguistics.

BRIN, SERGEY, & LAWRENCE PAGE. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks* 30.107–117.

Brown, L. D., T. T. Cai, & A. DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical Science* 16.101–117.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, & Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19.263–311.

Budanitsky, Alexander, & Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics* 32.13–47.

Buyko, Ekaterina, Katrin Tomanek, & Udo Hahn. 2007. Resolution of coordination ellipses in biological named entities using conditional random fields. 163–171.

Cafarella, Michael, Michele Banko, & Oren Etzioni. 2006. Relational Web search. Technical Report 2006-04-02, University of Washington, Department of Computer Science and Engineering.

Callison-Burch, Chris, Philipp Koehn, & Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

Calvo, Hiram, & Alexander Gelbukh. 2003. Improving prepositional phrase attachment disambiguation using the web as corpus. In *Progress in Pattern Recognition, Speech and Image Analysis: 8th Iberoamerican Congress on Pattern Recognition, CIARP 2003*, 604–610.

Cao, Yunbo, & Hang Li. 2002. Base noun phrase translation using web data and the EM algorithm. In *Proceedings of the 19th international conference on Computational linguistics*, 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Chantree, Francis, Adam Kilgarriff, Anne De Roeck, & Alistair Willis. 2005. Disambiguating coordinations using word distribution information. In *Proceedings of: RANLP 2005 (Recent Advances in Natural Language Processing)*, 144–151.

Charniak, Eugene, & Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 173–180, Morristown, NJ, USA. Association for Computational Linguistics.

Chklovski, Timothy, & Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 33–40.

Chomsky, Noam, & Morris Halle. 1991. *Sound pattern of English*. MIT Press, 2 edition.

——, Morris Halle, & Fred Lukoff. 1956. On accent and juncture in english. *In M. Halle, H.G. Lunt, H. McLean and C.H. van Schooneveld, eds. For Roman Jakobson: Essays on the occasion of his sixtieth birthday.* 65–80.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20.37–46.

——. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 2 edition.

COLLINS, MICHAEL. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics* 29.589–637.

——, & JAMES BROOKS. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of EMNLP*, 27–38.

DEVEREUX, BARRY, & FINTAN COSTELLO. 2006. Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. 184–189, Hillsdale, NJ. Erlbaum.

DOWNING, PAMELA. 1977. On the creation and use of english compound nouns. *Language* 810–842.

DUMAIS, SUSAN, MICHELE BANKO, ERIC BRILL, JIMMY LIN, & ANDREW NG. 2002. Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 291–298, New York, NY, USA. ACM Press.

DZHAMBAZOV, PETYR, YORDANKA PYSHEVA, TAMARA MAHROVA, & KINA IVANOVA. 1994. *Morfologiya russkogo yazyka. Prakticheskiy kurs. (in Russian)*. Veliko Turnovo, Bulgaria: Universitetsko izdatelstvo "Sv. sv. Kiril i Metodiy".

ETZIONI, OREN, MICHAEL CAFARELLA, DOUG DOWNEY, ANA-MARIA POPESCU, TAL SHAKED, STEPHEN SODERLAND, DANIEL S. WELD, & ALEXANDER YATES. 2005.

Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence* 165.91–134.

Fellbaum, Christiane (ed.) 1998. *Wordnet: An electronic lexical database.* MIT Press.

Fillmore, Charles J. 1968. Universals in linguistic theory. 1–88, New York. Holt, Rinehart, and Winston.

——. 1982. Frame semantics. 111–137, Seoul. Hanshin Publishing Co.

Finin, Timothy, 1980. *The semantic interpretation of compound nominals.* Urbana, Illinois: University of Illinois dissertation.

Fleiss, Joseph L. 1981. *Statistical methods for rates and proportions.* New York: John Wiley & Sons, Inc., 2 edition.

Gagné, Christina L. 2002. Lexical and relational influences on the processing of novel compounds. *Brain and Language* 81.723–735.

Girju, R., D. Moldovan, M. Tatu, & D. Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions* 4.479–496.

Girju, Roxana. 2006. Out-of-context noun phrase semantic interpretation with cross-linguistic evidence. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 268–276, New York, NY, USA. ACM.

——. 2007a. Experiments with an annotation scheme for a knowledge-rich noun phrase interpretation system. In *Proceedings of the Linguistic Annotation Workshop*, 168–175, Prague, Czech Republic. Association for Computational Linguistics.

——. 2007b. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 568–575, Prague, Czech Republic. Association for Computational Linguistics.

——, Ana-Maria Giuglea, Marian Olteanu, Ovidiu Fortu, Orest Bolohan, & Dan Moldovan. 2004. Support vector machines applied to the classification of semantic relations in nominalized noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, 68–75, Boston, MA.

——, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, & Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 13–18, Prague, Czech Republic. Association for Computational Linguistics.

Gochev, Gocho, Siyka Gocheva, & Todor Slavkov. 1987. *Prakticheskoe posobie po russkoy orfografii (in Russian)*. Sofia, Bulgaria: Narodna prosveta.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Miriam. 1999. An unsupervised model for statistically determining coordinate phrase attachment. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 610–614, Morristown, NJ, USA. Association for Computational Linguistics.

Grefenstette, Gregory. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Translating and the Computer 21: Proceedings of the*

21st International Conference on Translating and the Computer.

Grishman, Ralph, & John Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th conference on Computational linguistics*, 742–747, Morristown, NJ, USA. Association for Computational Linguistics.

Hindle, Donald, & Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics* 19.103–120.

Hogan, Deirdre. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 680–687, Prague, Czech Republic. Association for Computational Linguistics.

Idan Szpektor, Hristo Tanev, Ido Dagan, & Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 41–48.

Jackendoff, Ray. 1975. Morphological and semantic regularities in the lexicon. *Language* 51.639–671.

——. 1983. *Semantics and cognition*. Cambridge, MA: MIT Press.

Katz, Jerry, & Jerry Fodor. 1963. The structure of a semantic theory. *Language* 170–210.

Kauchak, David, & Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the main conference on Human Language Technology Conference of* *the North American Chapter of the Association of Computational Linguistics*, 455–462, Morristown, NJ, USA. Association for Computational Linguistics.

Keller, Frank, & Mirella Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29.459–484.

Kilgariff, Adam, & Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29.333–347.

Kilgarriff, Adam. 2007. Googleology is bad science. *Comput. Linguist.* 33.147–151.

Kim, Su Nam, & Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 491–498, Morristown, NJ, USA. Association for Computational Linguistics.

——, & Timothy Baldwin. 2007a. Disambiguating noun compounds. In *Proceedings of the 22th Conference on Artificial Intelligence*, 901–906.

——, & Timothy Baldwin. 2007b. Interpreting noun compounds via bootstrapping and sense collocation. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 129–136.

Klein, Dan, & Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 423–430, Morristown, NJ, USA. Association for Computational Linguistics.

KOEHN, PHILIPP. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of Sixth Conference of the Association for Machine Translation in the Americas (AMTA) 2004*, 115–124.

——. 2005. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of the X MT Summit*, 79–86, Phuket, Thailand.

——, & KEVIN KNIGHT. 2003. Feature-rich statistical translation of noun phrases. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

KUROHASHI, SADAO, & MAKOTO NAGAO. 1992. Dynamic programming method for analyzing conjunctive structures in Japanese. In *Proceedings of the 14th conference on Computational linguistics*, 170–176, Morristown, NJ, USA. Association for Computational Linguistics.

LAPATA, MARIA. 2002. The disambiguation of nominalizations. *Computational Linguistics* 28.357–388.

LAPATA, MIRELLA, & FRANK KELLER. 2004. The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. In *Proceedings of HLT-NAACL*, 121–128, Boston.

——, & FRANK KELLER. 2005. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.* 2.3.

LAUER, MARK, 1995. *Designing statistical language learners: Experiments on noun compounds*. Department of Computing Macquarie University NSW 2109 Australia dissertation.

Levi, Judith. 1978. *The syntax and semantics of complex nominals.* New York: Academic Press.

Li, Hang. 2002. Word clustering and disambiguation based on co-occurrence data. *Natural Language Engineering* 8.25–42.

——, & Naoki Abe. 1998. Generalizing case frames using a thesaurus and the mdl principle. *Computational Linguistics* 24.217–244.

——, & Naoki Abe. 1999. Learning dependencies between case frame slots. *Computational Linguistics* 25.283–291.

Liberman, Mark, & Richard Sproat. 1992. The stress and structure of modified noun phrases in english. *Lexical Matters* 131–181.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

——, & Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering* 7.343–360.

Macleod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett, & Ruth Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proceedings of EURALEX'98*, Liege, Belgium.

Marcus, Mitchell. 1980. *A theory of syntactic recognition for natural language.* MIT Press.

——, Beatrice Santorini, & Mary Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19.313–330.

Mihalcea, Rada, & Dan I. Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 152–158, Morristown, NJ, USA. Association for Computational Linguistics.

Minnen, G., J. Carroll, & D. Pearce. 2001. Applied morphological processing of english. *Natural Language Engineering* 7.207–223.

Modjeska, Natalia N., Katja Markert, & Malvina Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 176–183, Morristown, NJ, USA. Association for Computational Linguistics.

Moldovan, Dan, Adriana Badulescu, Marta Tatu, Daniel Antohe, & Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, 60–67, Boston, MA.

Nagata, Masaaki, Teruka Saito, & Kenji Suzuki. 2001. Using the web as a bilingual dictionary. In *Proceedings of the workshop on Data-driven methods in machine translation*, 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Nakov, Preslav, & Marti Hearst. 2005a. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 17–24, Ann Arbor, Michigan. Association for Computational Linguistics.

——, & Marti Hearst. 2005b. A study of using search engine page hits as a proxy for n-gram frequencies. In *Proceedings of RANLP'2005*, 347–353, Borovets, Bulgaria.

——, & Marti Hearst. 2005c. Using the web as an implicit training set: application to structural ambiguity resolution. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 835–842, Morristown, NJ, USA. Association for Computational Linguistics.

——, & Marti Hearst. 2007. UCB system description for the WMT 2007 shared task. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 212–215, Prague, Czech Republic. Association for Computational Linguistics.

——, & Marti A. Hearst. 2006. Using verbs to characterize noun-noun relations. In *AIMSA*, ed. by Jerome Euzenat & John Domingue, volume 4183 of *Lecture Notes in Computer Science*, 233–244. Springer.

——, Ariel Schwartz, & Marti Hearst. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of SIGIR'04 Workshop on Search and Discovery in Bioinformatics*, 81–88, Sheffield, UK.

——, Ariel Schwartz, Brian Wolf, & Marti Hearst. 2005a. Scaling up BioNLP: Application of a text annotation architecture to noun compound bracketing. In *Proceedings of SIG BioLINK*.

——, Ariel Schwartz, Brian Wolf, & Marti Hearst. 2005b. Supporting annotation layers for natural language processing. In *ACL '05: Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, 65–68, Morristown, NJ, USA. Association for Computational Linguistics.

NULTY, PAUL. 2007. Semantic classification of noun phrases using web counts and learning algorithms. In *Proceedings of the ACL 2007 Student Research Workshop*, 79–84, Prague, Czech Republic. Association for Computational Linguistics.

OCH, FRANZ JOSEF. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

——, & HERMANN NEY. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29.19–51.

——, & HERMANN NEY. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics* 30.417–449.

OKUMURA, AKITOSHI, & KAZUNORI MURAKI. 1994. Symmetric pattern matching analysis for english coordinate structures. In *Proceedings of the fourth conference on Applied natural language processing*, 41–46, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

OLTEANU, MARIAN, & DAN MOLDOVAN. 2005. PP-attachment disambiguation using large context. In *HLT'05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 273–280, Morristown, NJ, USA. Association for Computational Linguistics.

OSENOVA, PETYA, & KIRIL SIMOV. 2007. *Formalna gramatika na bylgarskiya ezik (in Bulgarian)*. Sofia, Bulgaria: IPP, BAS.

Padó, Sebastian, & Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33.161–199.

Pang, Bo, Kevin Knight, & Daniel Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 102–109, Morristown, NJ, USA. Association for Computational Linguistics.

Pantel, Patrick, & Dekang Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 101–108, Morristown, NJ, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, & Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Petruck, Miriam. 1996. Frame semantics. In *Handbook of Pragmatics*, ed. by J. Verschueren, J-O stman, J. Blommaert, & C. Bulcaen, 1–13. John Benjamins.

Porter, Martin. 1980. An algorithm for suffix stripping. *Program* 14.130–137.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: MIT Press.

——, Peter Anick, & Sabine Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics* 19.331–358.

QUINLAN, J. R., & R. L. RIVEST. 1989. Inferring decision trees using the minimum description length principle. *Information and Computation* 80.227–248.

QUINLAN, J. ROSS. 1993. *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

QUIRK, RANDOLPH, SIDNEY GREENBAUM, GEOFFRY LEECH, & JAN SVARTVIK. 1985. *A comprehensive grammar of the english language*. Longman.

RATNAPARKHI, A., S. ROUKOS, & R. WARD. 1994a. A maximum entropy model for parsing. In *Proceedings of the International Conference on Spoken Language Processing*, 803–806.

RATNAPARKHI, ADWAIT. 1998. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of COLING-ACL*, volume 2, 1079–1085.

——, JEFF REYNAR, & SALIM ROUKOS. 1994b. A maximum entropy model for prepositional phrase attachment. In *HLT '94: Proceedings of the workshop on Human Language Technology*, 250–255, Morristown, NJ, USA. Association for Computational Linguistics.

RESNIK, PHILIP, 1993. *Selection and information: a class-based approach to lexical relationships*. University of Pennsylvania, UMI Order No. GAX94-13894 dissertation.

——. 1999a. Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 527–534, Morristown, NJ, USA. Association for Computational Linguistics.

——. 1999b. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. (JAIR)* 11.95–130.

Rissanen, Jorma. 1978. Modelling by shortest data description. *Automatica* 14.465–471.

Romalo, Valeria Guţu, & *al.* 2005. *Gramatica limbii române.* Bucureşti, Romania: Editura Academiei Române.

Rosario, Barbara, & Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, ed. by Lillian Lee & Donna Harman, 82–90.

——, Marti A. Hearst, & Charles Fillmore. 2002. The descent of hierarchy, and selection in relational semantics. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 247–254, Morristown, NJ, USA. Association for Computational Linguistics.

Rozental', Ditmar. 1967. *Russkiy yazyk. Posobie dlya postupayushchih v vuzy (in Russian).* Moscow, USSR: Izdatel'stvo MGU.

——. 1970. *Voprosy russkogo pravopisaniya (in Russian).* Moscow, USSR: Izdatel'stvo MGU.

——. 1977. *Prakticheskaya stilistika russkogo yazyka (in Russian).* Moscow, USSR: Vysshaya shkola.

Ruge, Gerda. 1992. Experiment on linguistically-based term associations. *Inf. Process. Manage.* 28.317–332.

Rus, Vasile, Dan I. Moldovan, & Orest Bolohan. 2002. Bracketing compound nouns for logic form derivation. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, 198–202. AAAI Press.

Russo-Lassner, Grazia, Jimmy Lin, & Philip Resnik. 2005. A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland.

Saeed, John. 2003. *Semantics*. Blackwell, 2 edition.

Santamaría, Celina, Julio Gonzalo, & Felisa Verdejo. 2003. Automatic association of web directories with word senses. *Computational Linguistics* 29.485–502.

Séaghdha, Diarmuid Ó, & Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions*, 57–64, Prague, Czech Republic. Association for Computational Linguistics.

Shimbo, Masashi, & Kazuo Hara. 2007. A discriminative learning model for coordinate conjunctions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 610–619.

Shinyama, Yusuke, Satoshi Sekine, & Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the second international conference* *on Human Language Technology Research*, 313–318, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shinzato, Keiji, & Kentaro Torisawa. 2004. Acquiring hyponymy relations from web documents. In *HLT-NAACL*, 73–80.

Siegel, Sidney, & N. John Castellan Jr. 1988. *Non parametric statistics for the behavioral sciences*. New York: McGraw-Hill, 2 edition.

Soricut, Radu, & Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. *Information Retrieval* 9.191–206.

Stetina, Jiri, & Makoto Nagao. 1997. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of WVLC*, 66–80.

Tanaka, Takaaki, & Timothy Baldwin. 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

Tatu, Marta, & Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 371–378, Morristown, NJ, USA. Association for Computational Linguistics.

Toutanova, K., & C. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Confer-*

ence on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), 63–71.

Toutanova, Kristina, Christopher D. Manning, & Andrew Y. Ng. 2004. Learning random walk models for inducing word dependency distributions. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, p. 103, New York, NY, USA. ACM Press.

Trask, Robert. 1993. *A dictionary of grammatical terms in linguistics*. New York: Routledge.

Turney, Peter. 2005. Measuring semantic similarity by latent relational analysis. In *In Proceedings Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 1136–1141.

——, & Michael Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning Journal* 60.251–278.

——, Michael Littman, Jeffrey Bigham, & Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *In Proceedings of Recent Advances in Natural Language Processing (RANLP-03)*, 482–489.

Turney, Peter D. 2006a. Expressing implicit semantic relations without supervision. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, 313–320, Morristown, NJ, USA. Association for Computational Linguistics.

——. 2006b. Similarity of semantic relations. *Computational Linguistics* 32.379–416.

VADAS, DAVID, & JAMES CURRAN. 2007. Adding noun phrase structure to the penn treebank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 240–247, Prague, Czech Republic. Association for Computational Linguistics.

VANDERWENDE, LUCY. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of COLING-94*, 782–788.

VANSCHOENWINKEL, BRAM, & BERNARD MANDERICK. 2003. A weighted polynomial information gain kernel for resolving prepositional phrase attachment ambiguities with support vector machines. In *IJCAI*, ed. by Georg Gottlob & Toby Walsh, 133–140. Morgan Kaufmann.

VEALE, TONY. 2004. Wordnet sits the s.a.t. - a knowledge-based approach to lexical analogy. In *Proceedings of ECAI-2004*, 606–612.

VÉRONIS, JEAN. 2005a. Google: A snapshot of the update. *Jean Veronis' blog: http://aixtal.blogspot.com/2005/03/google-snapshot-of-update.html* .

——. 2005b. Web: Google adjusts its counts. *Jean Veronis' blog: http://aixtal.blogspot.com/2005/03/web-google-adjusts-its-counts.html* .

——. 2005c. Web: Google's missing pages: mystery solved? *Jean Veronis' blog: http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html* .

——. 2005d. Web: MSN cheating too? *Jean Veronis' blog: http://aixtal.blogspot.com/2005/02/web-msn-cheating-too.html* .

——. 2005e. Web: Yahoo doubles its counts! *Jean Veronis' blog: http://aixtal.blogspot.com/2005/03/web-yahoo-doubles-its-counts.html* .

Villavicencio, Aline, Francis Bond, Anna Korhonen, & Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language* 19.365–377.

Volk, Martin. 2000. Scaling up. using the WWW to resolve PP attachment ambiguities. In *KONVENS 2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), 6. ITG-Fachtagung "Sprachkommunikation"*, 151–155, Berlin, Germany, Germany. VDE-Verlag GmbH.

——. 2001. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics*, 601–606.

Wagner, Andreas, 2005. *Learning thematic role relations for lexical semantic nets*. Neuphilologischen Fakultät der Universität Tübingen, Germany dissertation.

Warren, Beatrice. 1978. Semantic patterns of noun-noun compounds. In *Gothenburg Studies in English 41, Goteburg, Acta Universtatis Gothoburgensis*.

Way, Andy, & Nano Gough. 2003. wEBMT: developing and validating an example-based machine translation system using the world wide web. *Computational Linguistics* 29.421–457.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 209–212.

Zahariev, Manuel, 2004. *A (acronyms)*. School of Computing Science, Simon Fraser University, USA dissertation.

Zavrel, Jakub, Walter Daelemans, & Jorn Veenstra. 1997. Resolving PP attachment ambiguities with memory-based learning. In *CoNLL97: Computational natural language learning*, ed. by T. Mark Ellison, 136–144. Somerset, New Jersey: Association for Computational Linguistics.

Zhai, Chengxiang. 1997. Fast statistical parsing of noun phrases for document indexing. In *Proceedings of the fifth conference on Applied natural language processing*, 312–319, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, Ying, & Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *TMI '04: 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, 4–6.

Zhao, Shaojun, & Dekang Lin. 2004. A nearest-neighbor method for resolving PP-attachment ambiguity. In *IJCNLP*, 545–554.

Zhou, Liang, Chin-Yew Lin, & Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 77–84, Sydney, Australia. Association for Computational Linguistics.

Zhu, Xiaojin, & Ronald Rosenfeld. 2001. Improving trigram language modeling with the world wide web. In *Proceedings of ICASSP*, I:533–536.

Appendix A

Semantic Entailment

For a typical right-headed non-lexicalized noun-noun compound n_1n_2, it can be expected that the assertion "n_1n_2 is a type/kind of n_2" (e.g., "*lung cancer* is a kind/type of *cancer*") would be true, i.e. that n_1n_2 semantically entails n_2. Note that this is not true for all noun-noun compounds, e.g., *birdbrain* is not a kind/type of *brain*; it is a kind of person. Similarly, *vitamin D* is not a kind/type of *D*; it is a kind of *vitamin*.

For a three-word right-headed non-lexicalized noun compound $n_1n_2n_3$, it can be expected that the assertion "$n_1n_2n_3$ is a type/kind of n_3" would be true, e.g., "*lung cancer drug* is a kind/type of *drug*". If the compound is left-bracketed $(n_1n_2)n_3$, it can be further expected that the assertion "$n_1n_2n_3$ is a type/kind of n_2n_3" would be true as well, e.g., "*lung cancer drug* is a kind/type of *cancer drug*". If the compound is right-bracketed $n_1(n_2n_3)$, then it can be expected that the following two assertions are true "$n_1n_2n_3$ is a type/kind of n_2n_3" and "$n_1n_2n_3$ is a type/kind of n_1n_3", e.g., "*oxydant wrinkle treatment* is a kind/type of *wrinkle treatment*" and "*oxydant wrinkle treatment* is a kind/type of *oxydant treatment*".

$\mathbf{n_1 n_2 n_3}$	$\mathbf{n_1 n_3}$	"⇒"	$\mathbf{n_2 n_3}$	"⇒"	**Brack.**
army ant behavior	*army behavior*	no	*ant behavior*	yes	left
health care reform	*health reform*	no	*care reform*	**NO**	left
heart rate variability	*heart variability*	**YES**	*rate variability*	yes	left
lung cancer doctor	*lung doctor*	**YES**	*cancer doctor*	yes	left
lung cancer physician	*lung physician*	**YES**	*cancer physician*	yes	left
lung cancer patient	*lung patient*	**YES**	*cancer patient*	yes	left
lung cancer survivor	*lung survivor*	no	*cancer survivor*	yes	left
lung cancer treatment	*lung treatment*	**YES**	*cancer treatment*	yes	left
lung cancer drug	*lung drug*	no	*cancer drug*	yes	left
science fiction satire	*science satire*	no	*fiction satire*	**NO**	left
science fiction writer	*science writer*	no	*fiction writer*	**NO**	left
science fiction novels	*science novels*	no	*fiction novels*	**NO**	left
US army forces	*US forces*	**YES**	*army forces*	yes	left
vitamin D deficiency	*vitamin deficiency*	**YES**	*D deficiency*	**NO**	left
oxydant wrinkle treatment	*oxydant treatment*	yes	*wrinkle treatment*	yes	right
space science fiction	*space fiction*	**NO**	*science fiction*	yes	right
brain stem cell	*brain cell*	yes	*stem cell*	yes	both

Table A.1: **Semantic entailment and noun compound bracketing.** The entailments that are inconsistent with the bracketing are shown in bold.

In other words, given a three-word right-headed non-lexicalized noun compound $(n_1 n_2) n_3$, it is expected to entail n_3 and $n_2 n_3$, but not $n_1 n_3$, if left-bracketed, and all three n_3, $n_2 n_3$, and $n_1 n_3$, if right-bracketed. Therefore, noun compound bracketing can be used to suggest semantic entailment. Unfortunately, there is no easy way to distinguish left-headed from right-headed compounds, and it is even harder to distinguish between lexicalized and non-lexicalized compounds, the boundary between which is not well-established. The matter is further complicated due to nominalization, metaphor, metonymy, contextual dependency, world knowledge, and pragmatics.

Table A.1 shows sample three-word noun compounds, the corresponding bracketing and whether $n_1 n_3$ and $n_2 n_3$ are entailed or not. The entailments that are inconsistent with the bracketing are shown in bold. For some of them, it is easy to come up an explanation:

1. **Lexicalization/morphology**. As I mentioned in section 2.2, noun compounds lie on the boundary between lexicalization/morphology and syntax. While some are completely lexicalized (e.g. *birdbrain*), and other are purely compositional (e.g., *ant behavior*), the majority occupy the continuum in between. For example, *science fiction* looks lexicalized to a higher degree compared to *lung cancer* and *heart rate*, which in turn are more lexicalized compared to *cancer drug*. In the process of lexicalization, noun-noun compounds become less and less compositional, which could explain the *science fiction* examples: *science fiction* is not a kind of *fiction*. The same explanation can be applied to the *health care* example, which is probably even more lexicalized as it is often written as a single word *healthcare*: while its meaning does not depart too far from a pure compositional interpretation, it is non-compositional enough to prevent *health care reform* from entailing *care reform*.

2. **Individual Nouns' Semantics**. The fact that *lung cancer physician* implies *lung physician* cannot be explained by the relatively low degree of lexicalization of *lung cancer* alone. The noun compound is left bracketed and therefore this entailment should not be true. It looks like the semantics of the individual nouns and world knowledge come into play, e.g., *physician* is likely to be modified by organs.

3. **Nominalization**. The really problematic noun compounds are *lung cancer patient* and *lung cancer survivor*. While *patient* and *survivor* have a very similar meaning in the context of these compounds, *lung patient* is implied, but *lung survivor* is not. One possible explanation is that *survivor* is a nominalization of the verb *survive*, which likes *cancer*, and therefore *lung cancer*, as a direct object, but does not like *lung*.

4. **Left Head**. The entailments are completely reversed for *vitamin D deficiency*, which contains the left-headed noun compound *vitamin D* as a modifier: the interpretation of the noun compound changes completely when the head precedes the modifier.

5. **Metonymy**. Due to metonymy, in the case of *US army forces*, all possible subsequences are entailed: not only *US forces*, *army forces* and *forces*, but also *US army*, *army* and, in some contexts, even just *US*.

Appendix B

Phrase Variability in Textual Entailment

Here I present a short study of some examples from the development dataset of the Second Pascal Recognizing Textual Entailment (RTE2) Challenge.[1] The challenge addresses a generic semantic inference task needed by many natural language processing applications, including Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and (multi-)document summarization (SUM). Given two textual fragments, a text T and a hypothesys H, the goal is to recognize whether the meaning of H is entailed (can be inferred) from T:

> "*We say that T entails H if, typically, a human reading T would infer that H is most likely true. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge.*" (RTE2 task definition)

[1] `http://www.pascal-network.org/Challenges/RTE2`

Below I list some examples where T does entail H in order to illustrate different kinds of phrase variability problems that a successful inference procedure might need to solve. The target phrases are underlined.

B.1 Prepositional Paraphrase

Here a noun compound has to be matched to a prepositional paraphrase.

"*paper costs*" $\Rightarrow$ "*cost of paper*"

```
<pair id="503" entailment="YES" task="IR">
<t>Newspapers choke on rising paper costs and falling revenue.</t>
<h>The cost of paper is rising.</h>
</pair>
```

B.2 Prepositional Paraphrase & Genitive Noun Compound

Here a genitive noun compound needs to be matched to a prepositional paraphrase.

"*Basra's governor*" $\Longrightarrow$ "*governor of Basra*"

```
<pair id="337" entailment="YES" task="SUM">
<t>Basra's governor said he would not cooperate with British troops until
there was an apology for a raid to free two UK soldiers.</t>
<h>The governor of Basra will not work with British troops until
there is an apology for a raid to free two UK soldiers.</h>
</pair>
```

B.3 Verbal & Prepositional Paraphrases

Here we have two different paraphrases, one verbal and one prepositional, of the same noun compound *WTO Geneva headquarters*.

"*Geneva headquarters of the WTO*" ⇒ "*WTO headquarters are located in Geneva*"

```
<pair id="284" entailment="YES" task="QA">
<t>While preliminary work goes on at the Geneva headquarters of the WTO,
with members providing input, key decisions are taken at the ministerial
meetings.</t>
<h>The WTO headquarters are located in Geneva.</h>
</pair>
```

B.4 Nominalization

Here we have two nouns representing different nominalizations referring to the same process.

"*legalizing drugs*" ⇒ "*drug legalization*"

```
<pair id="103" entailment="YES" task="IR">
<t>This paper describes American alcohol use, the temperance movement,
Prohibition, and the War on Drugs and explains how legalizing drugs would
reduce crime and public health problems.</t>
<h>Drug legalization has benefits.</h>
</pair>
```

B.5 Bracketing

Here we have an entailment which extracts a component consistent with a left bracketing.

"[*breast cancer*] *patients*" ⇒ "[*breast cancer*]"

```
<pair id="177" entailment="YES" task="SUM">
<t>Herceptin was already approved to treat the sickest
breast cancer patients, and the company said, Monday,
it will discuss with federal regulators the possibility
of prescribing the drug for more breast cancer patients.</t>
<h>Herceptin can be used to treat breast cancer.</h>
</pair>
```

Here is a longer example where the extracted component is consistent with a left bracketing.

"[*ocean* [*remote sensing*]] *science*" ⇒ "[*ocean* [*remote sensing*]]"

```
<pair id="155" entailment="YES" task="IR">
<t>Scientists and engineers in the APL Space Department have
contributed to ocean remote sensing science and technology
for more than a quarter century.</t>
<h>Ocean remote sensing is developed.</h>
</pair>
```

B.6 Bracketing & Paraphrase

Here we have a noun compound which is paraphrased using a preposition in a way that is consistent with left bracketing.

"[*ivory trade*] *ban*" ⇒ "*ban on* [*ivory trade*]"

```
<pair id="117" entailment="NO" task="IR">
<t>The release of its report led to calls for a complete ivory trade ban,
and at the seventh conference in 1989, the African Elephant was moved to
appendix one of the treaty.</t>
<h>The ban on ivory trade has been effective in protecting the elephant
from extinction.</h>
</pair>
```

B.7 Synonymy

Here we have two noun compounds with the same modifiers and synonymous heads.

"*marine plants*" ⇒ "*marine vegetation*"

```
<pair id="65" entailment="YES" task="IR">
<t>A number of marine plants are harvested commercially
in Nova Scotia.</t>
<h>Marine vegetation is harvested.</h>
</pair>
```

B.8 Hyponymy & Prepositional Paraphrase

Here we have a prepositional paraphrase and a noun compound with the same modifiers and heads that are in a hyponymy relation[2]: *marijuana* is a kind of *drug*.

"*legalization of marijuana*" ⇒ "*drug legalization*"

```
<pair id="363" entailment="YES" task="IR">
<t>One economic study will not be the basis of Canada's public policy decisions,
but Easton's research does conclusively show that there are economic benefits
in the legalization of marijuana.</t>
<h>Drug legalization has benefits.</h>
</pair>
```

B.9 Prepositional Paraphrase, Nominalization & Synonymy

Here the verb *enrich* has been substitued by its synonym *enhance*, which in turn has been nominalized. On the other hand, the verbs *enrich* and *feed* are good paraphrases for the noun compound *soil enhancers*, e.g., "*enhancers that enrich the soil*", and *fertilizer* is a contextual synonym of *enhancer*.

"*enriches and feeds the soil*" ⇒ "*soil enhancers*"

[2]Alternatively, we have paraphrases for the longer noun compounds *marijuana legalization benefits* and *drug legalization benefits*, i.e., "*benefits in the legalization of marijuana*" ⇒ "*drug legalization has benefits*".

```
<pair id="192" entailment="YES" task="IR">
<t>Organic fertilizer slowly enriches and feeds the soil.  Fast acting synthetic
fertilizers harm soil life.</t>
<h>Organic fertilizers are used as soil enhancers.</h>
</pair>
```

B.10 Nominalization, Bracketing & Synonymy

The example this time is Here the modificational past participle *damaged* becomes the new noun compound head: *injury*. One possible interpretation is that *damaged*, which is a form of the verb *to damage*, has been substituted by the synonymous verb *to injure*, which has been nominalized and has become the new noun compound head. Note however that the verb *to damage* does not have to be analyzed as a synonym of *to injure*; it is also a good candidate for a verbal paraphrase of *spinal cord injury*, e.g., as "*injury which damaged the spinal cord*". Note that while *damaged spinal cord* does entail *spinal cord injury*, the entailment does not hold for the overall example.

"*damaged* [*spinal cord*]" $\Rightarrow$ "[*spinal cord*] *injury*"

```
<pair id="261" entailment="NO" task="SUM">
<t>The new work went an extra step, suggesting that the connections that the stem
cells form to help bridge the damaged spinal cord, are key to recovery.</t>
<h>The experiment, reported Monday, isn't the first to show that stem cells offer
tantalizing hope for spinal cord injury.</h>
</pair>
```

B.11 Discussion

As I have shown, solving many, if not all, of the above textual entailments might require understanding the syntax and/or the semantics of the involved noun compounds. Some of the examples require checking whether a particular verb or preposition is acceptable as a paraphrase for a given noun compound, other call for judging on semantic entailment after word dropping from a noun compound. Additional factors come into play as well, e.g., genitives, inflectional and derivational morphological alternations, lexical relations like synonymy and hyponymy.

An important observation to make is that noun compound interpretation techniques can naturally extend to cases where noun compounds are formally not involved. For example, checking whether "*Geneva headquarters of the WTO*" could textually entail "*WTO headquarters are located in Geneva*" can be performed by understanding the semantics of the noun compound "*WTO Geneva headquarters*" in terms of possible prepositional and verbal paraphrases.

Appendix C

Biomedical Dataset for Noun Compound Bracketing

C.1 Extracting the Most Frequent Noun Compounds from MEDLINE

The three-word noun compounds for the *Biomedical dataset* for noun compound bracketing are extracted from a collection of 1.4 million MEDLINE abstracts (citations between 1994 and 2003), which constitutes about 8.23% of the total 17 million abstracts in MEDLINE.[1]

The abstracts have been sentence split, tokenized, POS tagged and shallow parsed using the Open NLP tools.[2] The resulting linguistic annotations – sentences, tokens, POS, and shallow parses – were indexed by a specialized system architecture developed

[1] `http://www.ncbi.nlm.nih.gov/sites/entrez`
[2] `http://opennlp.sourceforge.net/`

as part of the *Biotext project*[3], which supports queries over layers of annotation on natural language text. The system allows for both hierarchical and overlapping layers and for querying at multiple levels of description. The implementation is built on top of a standard relational database management system (RDBMS), and, by using carefully constructed indexes, it can execute complex queries efficiently. More information about the system[4] and its query language, the *Layered Query Language* (*LQL*), can be found in (Nakov *et al.* 2005a; Nakov *et al.* 2005b). A detailed description of the language syntax, a broader example-based introduction and an online demo can be found online at `http://biotext.berkeley.edu/lql/`.

Using the LQL query language, I extracted all sequences of three nouns falling in the last three positions of a noun phrase (NP) found in the shallow parse. If the NP contained other nouns, the sequence was discarded. This allows for noun compounds that are modified by adjectives, determiners, and so on, but prevents extracting three-word noun compounds that are part of longer noun compounds. Details follow below.

In order to simplify the explanations, I start with a simple LQL query and I refine it in two additional steps in order to obtain the final version that I use to extract the three-word noun compounds for the *Biomedical dataset*. Consider the following query:

The new query contains an assertion that a layer does not exist, via the negation operator ("`!`"). The outer brackets indicate that gaps (intermediate words) are allowed between any layers, thus disallowing the non-existing layer anywhere before the sequence of three nouns (not just when it immediately precedes the first noun). One more pair of

[3]`http://biotext.berkeley.edu`

[4]The LQL query language and the architecture have been developed by Ariel Schwartz, Brian Wolf, Gaurav Bhalotia, and myself under the supervision of Prof. Marti Hearst. Rowena Luk and Archana Ganapathi contributed to the design.

brackets counteracts ALLOW GAPS and keeps the three internal nouns adjacent.

Query (1):

```
FROM
   [layer='shallow_parse' && tag_name='NP'
     ^ [layer='pos' && tag_name="noun"] AS n1
       [layer='pos' && tag_name="noun"] AS n2
       [layer='pos' && tag_name="noun"] AS n3 $
   ]
SELECT n1.content, n2.content, n3.content
```

The above LQL query looks for a noun phrase from the shallow-parse layer, containing exactly three nouns from the POS layer. Each layer in an LQL query is enclosed in brackets and is optionally followed by a binding statement. Layers have names, e.g., `pos` and `shallow_parse`, which determine a possible set of types such as `NP` for `shallow_parse` and `NN` for `pos`. There are macros for groups of tags, such as `noun`, which refers to all parts of speech which are nouns in the Penn Treebank. Single quotes are used for exact matches, and double quotes for case insensitive matches and macros.

Enclosure of one layer within another indicates that the outer spans the inner. By default, layers' contents are adjacent, but the keywords `ALLOW GAPS` can change this default. LQL uses the UNIX-style delimiters `^` and `$` in order to constrain the results so that no other words can precede or follow the three nouns within the NP. Finally, the `select` statement specifies that the contents of the *compound* (i.e. the text span of the compound) are to be returned.

Query (1) is case sensitive. While this is useful in the biomedical domain, where the capitalization can distinguish between different genes or proteins, to get frequency counts we want normalization. *Query (2)* below converts the results to lowercase by using the corresponding SQL function.

Query (2) encloses the LQL inside an SQL block and then uses SQL aggregation functions to produce a sorted list (in descending order) of the noun compounds and their corresponding frequencies. Note the two `SELECT` statements – the first one belongs to SQL, and the second one to LQL. This query will extract three-word noun compounds, provided that the POS tagging and the shallow-parse annotations are correct. However, since it requires that the NP consist of exactly three nouns (without preceding adjectives, determiners etc.), it will miss some three-word noun compounds that are preceded by modifiers and determiners (such as *the*, *a*, *this*).

Query (2):

```
SELECT
    LOWER(n1.content) lc1,
    LOWER(n2.content) lc2,
    LOWER(n3.content) lc3,
    COUNT(*) AS freq
FROM
  BEGIN_LQL
     FROM
        [layer='shallow_parse' && tag_name='NP'
          ^ [layer='pos' && tag_name="noun"] AS n1
            [layer='pos' && tag_name="noun"] AS n2
            [layer='pos' && tag_name="noun"] AS n3 $
        ]
     SELECT n1.content, n2.content, n3.content
  END_LQL
GROUP BY lc1, lc2, lc3
ORDER BY freq DESC
```

Query (3) below asserts that the nouns should occupy the three last positions in the NP and disallows other nouns within the same NP, but allows for other parts of speech. Note the trade-off between query complexity and amount of control afforded by the language.

This new query contains an assertion that a layer does not exist, via the negation operator ("`!`"). The outer brackets indicate that gaps (intermediate words) are allowed between any layers, thus disallowing the non-existing layer anywhere before the sequence of three nouns (not just when it immediately precedes the first noun). One more pair of brackets counteracts `ALLOW GAPS` and keeps the three internal nouns adjacent.

Query (3):

```
SELECT LOWER(n1.content) lc1,
       LOWER(n2.content) lc2,
       LOWER(n3.content) lc3,
       COUNT(*) AS freq
FROM
  BEGIN_LQL
    FROM
      [layer='shallow_parse' && tag_name='NP'
        ^ ( { ALLOW GAPS }
            ![layer='pos' && tag_name="noun"]
             ( [layer='pos' && tag_name="noun"] AS n1
               [layer='pos' && tag_name="noun"] AS n2
               [layer='pos' && tag_name="noun"] AS n3 ) $
          ) $
      ]
    SELECT n1.content, n2.content, n3.content
  END_LQL
GROUP BY lc1, lc2, lc3
ORDER BY freq DESC
```

Query (3) returns a total of 418,678 distinct noun compounds with corresponding frequencies. I manually investigate the most frequent ones, removing those with errors in tokenization (e.g., containing words like *transplan* or *tation*), POS tagging (e.g., *acute lung injury*, where *acute* is wrongly tagged as a noun) or shallow parsing (e.g., *situ hybridization*, that misses *in*). I had to consider the first 843 examples in order to obtain 500 good ones, which suggests an extraction accuracy of 59%. This number is low for various reasons. First, the tokenizer handles dash-connected words as a single token (e.g., *factor-alpha*). Second, many tokens contain other special characters (e.g., *cd4+*), which I exclude since they cannot be used in a query against a search engine. Note that while the extraction process does not allow for exact repetitions, some examples are similar due to various kinds of variability:

- **spelling variation**
  - "*iron deficiency <u>anaemia</u>*" vs. "*iron deficiency <u>anemia</u>*";
  - "*motor <u>neuron</u> disease*" vs. "*motor <u>neurone</u> disease*", etc.
- **inflectional variation**
  - "*blood flow <u>velocity</u>*" vs. "*blood flow <u>velocities</u>*";
  - "*group b <u>streptococcus</u>*" vs. "*group b <u>streptococci</u>*", etc.
- **derivational variation**
  - "*bone marrow <u>transplantation</u>*" vs. "*bone marrow <u>transplants</u>*";
  - "*dna strand <u>breakage</u>*" vs. "*dna strand <u>breaks</u>*", etc.
- **disease/gene/protein/chemical/etc. name variation**
  - "*hepatitis <u>c</u> virus*" vs. "*hepatitis <u>e</u> virus*";
  - "*rna polymerase <u>ii</u>*" vs. "*rna polymerase <u>iii</u>*";
- **hyponymy**
  - "*colon <u>cancer</u> cells*" vs. "*colon <u>carcinoma</u> cells*", etc.

C.2 Extracting Statistics from MEDLINE instead of the Web

Here I describe the process of acquisition of n-gram and paraphrase frequency statistics for the *Biomedical dataset* from 1.4 million MEDLINE abstracts with suitable linguistic annotations using LQL.

C.2.1 Extracting n-gram Frequencies

For a three-word noun compound $w_1w_2w_3$, the *adjacency* and the *dependency* models need the corpus frequencies for w_1w_2, w_1w_3 and w_2w_3. For the purpose, I use a case insensitive LQL query that allows for inflections of the second word, which respects the sentence boundaries and requires both words to be nouns and to be the last two words in a noun phrase. The LQL query below shows an example instantiation for the bigram "*immunodeficiency virus*":

```
SELECT
    COUNT(*) AS freq
FROM
  BEGIN_LQL
    FROM
      [layer='shallow_parse' && tag_name='NP'
        [layer='pos' && tag_name="noun"
           && content="immunodeficiency"] AS word1
        [layer='pos' && tag_name="noun"
           && (content="virus" || content="viruses")] $
      ]
    SELECT word1.content, word2.content
  END_LQL
```

The query for the word frequencies is much simpler. It only requires the word to be a noun, allowing for inflections. Only the total number of occurrences is selected. The LQL query below shows an example instantiation for the word "*human*":

```
SELECT COUNT(*) AS freq
FROM
  BEGIN_LQL
    FROM
      [layer='pos' && tag_name="noun" &&
        content IN ("human", "humans")] AS word
    SELECT word.content
  END_LQL
```

C.2.2 Extracting Paraphrase Frequencies

As I have already explained in section 3.7, paraphrases are an important feature type for noun compound bracketing. The example LQL query below follows the general pattern "$[_{\text{NP}} \ldots\ w_2\ w_3]$ *PREP* $[_{\text{NP}} \ldots\ w_1]$", looking for right-predicting **prepositional paraphrases** for the noun compound *human immunodeficiency virus*:

```
SELECT LOWER(prep.content) lp,
       COUNT(*) AS freq
FROM
  BEGIN_LQL
    FROM
      [layer='sentence'
        [layer='shallow_parse' && tag_name='NP'
           [layer='pos' && tag_name="noun" &&
               content = "immunodeficiency"]
           [layer='pos' && tag_name="noun" &&
               content IN ("virus", "viruses")] $
        ]
        [layer='pos' && tag_name='IN'] AS prep
        [layer='shallow_parse' && tag_name='NP'
           [layer='pos' && tag_name="noun" &&
               content IN ("human", "humans") ] $
        ]
      ]
    SELECT prep.content
  END_LQL
GROUP BY lp
ORDER BY freq DESC
```

The query above requires that the three words be nouns and allows for inflections of w_1 and w_3, but not of w_2 since it modifies w_3 in the paraphrase. It also required w_1 and "w_2 w_3" be the last words in their corresponding NPs, and accepts any preposition between these NPs.

Verbal paraphrases and **copula paraphrases** can be handled in a similar manner, but have been excluded from my experiments since they are very infrequent because of the relatively small size of my subset of MEDLINE. As already mentioned in section 3.8.2, my corpus consists of about 1.4 million MEDLINE abstracts, each one being about 300 words long on the average, which means about 420 million indexed words in total. For comparison, *Google* indexes about eight billion pages; assuming each one contains 500 words on the average, this means about four trillion indexed words, which is about a million times bigger than my corpus. Since the *prepositional* paraphrases are much more frequent and since I add up the total number of left-predicting vs. right-predicting paraphrase occurrences, the *verbal* and the *copula* paraphrases are unlikely to change the overall bracketing prediction of the paraphrase model: the total sums would still be dominated by the *prepositional* paraphrases. In addition, in the Web experiments above, the *verbal* paraphrases used a limited set of verbs which act as prepositions (see section 3.7 for a full list and for additional details). Since the LQL query above allows for any preposition, it would return some of these verbal paraphrases anyway. Therefore, I chose not to make use any *copula* or *verbal* paraphrases, thus limiting myself to *prepositional* paraphrases only which makes my experiments more directly comparable to the above described Web-based ones.

C.3 The Annotated *Biomedical Dataset*

Below I list the 430 test examples (sorted by decreasing frequency) from the *Biomedical dataset* for noun compound bracketing, described in section 3.8.2.

Noun Compound			Bracketing	Freq.
polymerase	chain	reaction	right	5208
body	mass	index	left	2096
nerve	growth	factor	right	1918
tumor	necrosis	factor	right	1888
positron	emission	tomography	left	1447
protein	kinase	c	left	1399
transmission	electron	microscopy	left	1331
bone	mineral	density	right	1257
hepatitis	c	virus	left	1211
bone	marrow	transplantation	left	979
bone	marrow	cells	left	918
t	cell	activation	left	912
heart	rate	variability	left	845
health	care	reform	left	842
lymph	node	metastasis	left	755
cell	cycle	progression	left	749
calcium	channel	blockers	right	739
health	care	providers	left	700
lymph	node	metastases	left	670
health	care	workers	left	634
dna	sequence	analysis	left	615
phorbol	myristate	acetate	left	615
growth	factor	beta	left	613
rna	polymerase	ii	left	608
stem	cell	factor	left	589
male	wistar	rats	right	570
hepatitis	b	virus	left	560
type	iv	collagen	left	560
pneumocystis	carinii	pneumonia	left	546
tissue	plasminogen	activator	right	515
tumor	suppressor	genes	left	510
breast	cancer	cells	left	503
breast	cancer	patients	left	486
dna	strand	breaks	right	481
health	care	professionals	left	456

Noun Compound			Bracketing	Freq.
signal	transduction	pathways	left	450
hepatocyte	growth	factor	right	450
t	cell	proliferation	left	444
protein	tyrosine	kinases	right	410
class	ii	molecules	left	397
xenopus	laevis	oocytes	left	394
body	weight	gain	left	392
body	surface	area	left	382
gel	filtration	chromatography	left	377
type	ii	collagen	left	375
heart	transplant	recipients	left	374
protein	tyrosine	phosphorylation	right	374
lymph	node	involvement	left	360
health	care	costs	left	358
amino	acid	sequence	left	350
health	care	services	left	349
amino	acid	residues	left	342
gel	mobility	shift	right	342
type	ii	cells	left	334
type	ii	diabetes	left	333
patent	ductus	arteriosus	right	327
escherichia	coli	cells	left	323
blood	pressure	control	left	314
cell	surface	expression	left	312
heat	shock	protein	left	308
blood	flow	velocity	left	305
tumour	necrosis	factor	right	305
t	cell	responses	left	302
growth	factor	alpha	left	298
health	care	delivery	left	297
amino	acid	sequences	left	291
breast	cancer	risk	left	291
t	cell	clones	left	290
motor	neuron	disease	left	287
lung	cancer	patients	left	285
lymph	node	cells	left	281
cell	cycle	arrest	left	271
mhc	class	ii	right	270
mycobacterium	avium	complex	left	262
growth	hormone	deficiency	left	261
t	cell	development	left	261
world	war	ii	left	260

Noun Compound			Bracketing	Freq.
amino	acid	substitutions	left	259
skin	blood	flow	right	259
borderline	personality	disorder	right	246
receptor	tyrosine	kinases	right	246
growth	factor	receptors	left	243
cell	cycle	control	left	229
wheat	germ	agglutinin	left	229
map	kinase	activation	left	228
t	cell	subsets	left	228
bone	mineral	content	right	227
t	cell	lines	left	226
calf	thymus	dna	left	225
casein	kinase	ii	left	224
cell	cycle	regulation	left	223
liver	transplant	recipients	left	221
dna	flow	cytometry	right	218
blood	urea	nitrogen	right	216
rat	liver	mitochondria	left	215
heat	shock	proteins	left	214
langerhans	cell	histiocytosis	left	213
reporter	gene	expression	left	211
type	iii	collagen	left	211
sister	chromatid	exchanges	left	210
pars	plana	vitrectomy	left	206
iron	deficiency	anemia	left	199
tyrosine	kinase	activity	left	199
phase	ii	study	left	198
group	b	patients	left	195
adhesion	molecule	expression	left	193
unit	cell	dimensions	right	193
laser	doppler	flowmetry	left	191
growth	hormone	secretion	left	190
injection	drug	users	left	190
ion	exchange	chromatography	left	189
newcastle	disease	virus	left	188
hepatitis	b	vaccine	left	186
nmda	receptor	activation	left	186
bladder	outlet	obstruction	left	185
silicone	breast	implants	right	185
dna	polymerase	alpha	right	183
ammonium	sulfate	precipitation	left	182
hepatitis	b	vaccination	left	181

Noun Compound			Bracketing	Freq.
monte	carlo	simulations	left	181
b	cell	development	left	179
dna	adduct	formation	left	178
primer	extension	analysis	left	178
skin	prick	tests	left	177
t	cell	recognition	left	177
gel	permeation	chromatography	left	173
ornithine	decarboxylase	activity	left	173
tyrosine	kinase	inhibitors	left	172
lymph	node	dissection	left	171
stage	iv	disease	left	169
nmda	receptor	antagonists	left	168
t	helper	cells	right	168
b	cell	activation	left	166
motor	vehicle	accidents	left	166
memory	t	cells	right	166
lymph	node	status	left	165
type	vi	collagen	left	165
kidney	transplant	recipients	left	164
liver	function	tests	left	164
blood	pressure	regulation	left	163
nerve	conduction	studies	left	163
plasminogen	activator	inhibitor	left	163
growth	hormone	treatment	left	162
substance	use	disorders	left	162
amino	acid	analysis	left	159
electron	spin	resonance	left	159
plasmodium	falciparum	malaria	left	159
rnase	protection	assay	left	159
cam	kinase	ii	left	157
group	b	streptococci	left	157
map	kinase	activity	left	157
group	ii	patients	left	156
merkel	cell	carcinoma	left	156
blood	pressure	reduction	left	154
islet	cell	antibodies	left	154
dobutamine	stress	echocardiography	right	154
b	cell	differentiation	left	153
breast	cancer	mortality	left	153
health	care	resources	left	153
restriction	endonuclease	analysis	left	153
mast	cell	degranulation	left	152

Noun Compound			Bracketing	Freq.
tumor	cell	growth	left	151
type	ii	pneumocytes	left	151
guinea	pig	ileum	left	150
erythrocyte	sedimentation	rate	left	150
blood	pressure	variability	left	148
rna	blot	analysis	left	148
anion	exchange	chromatography	left	147
class	ii	antigens	left	147
phase	ii	trial	left	147
proton	pump	inhibitors	left	147
tandem	mass	spectrometry	right	147
rat	brain	synaptosomes	left	145
glutathione	peroxidase	activity	left	144
blood	gas	analysis	left	143
blood	pressure	levels	left	142
choline	acetyltransferase	activity	left	142
lung	transplant	recipients	left	142
lung	cancer	risk	left	140
dna	topoisomerase	ii	left	140
heparan	sulfate	proteoglycans	left	137
bone	marrow	involvement	left	136
calcium	channel	antagonists	left	136
cell	cycle	analysis	left	136
cell	surface	molecules	left	136
vitamin	d	deficiency	left	135
protein	kinase	activity	left	134
hepatitis	c	infection	left	133
superoxide	anion	production	left	133
tissue	culture	cells	left	133
skin	biopsy	specimens	left	132
protein	disulfide	isomerase	left	131
tumor	cell	invasion	left	131
angiotensin	ii	receptors	left	130
density	gradient	centrifugation	left	130
phosphoamino	acid	analysis	left	130
rous	sarcoma	virus	left	130
blood	pressure	measurement	left	128
lung	function	tests	left	128
hepatitis	b	infection	left	127
home	health	care	right	127
mouse	hepatitis	virus	right	127
protein	tyrosine	kinase	right	127

Noun Compound			Bracketing	Freq.
health	maintenance	organizations	left	125
mast	cell	activation	left	125
restriction	enzyme	digestion	left	125
t	cell	function	left	125
chlamydia	trachomatis	infection	left	124
health	services	research	left	123
world	health	organization	right	122
amino	acid	sequencing	left	121
rat	vas	deferens	right	121
health	care	organizations	left	120
lactate	dehydrogenase	release	left	119
phase	ii	studies	left	119
protein	kinase	inhibitors	left	119
helper	t	cells	right	119
gamma	knife	radiosurgery	left	118
health	care	utilization	left	118
plasma	membrane	vesicles	left	118
protein	synthesis	inhibitors	left	118
cytokine	mrna	expression	left	118
escherichia	coli	strains	left	117
muscle	blood	flow	right	117
bile	duct	stones	left	116
health	care	systems	left	116
hyaline	membrane	disease	left	116
plasma	protein	binding	right	116
dna	binding	activity	left	115
rotator	cuff	tears	left	114
blood	pressure	measurements	left	113
sex	steroid	hormones	right	113
colon	cancer	cells	left	112
size	exclusion	chromatography	left	112
dentate	granule	cells	right	112
melanoma	cell	lines	right	112
monte	carlo	simulation	left	111
nerve	conduction	velocity	left	111
bundle	branch	block	left	110
germ	cell	tumors	left	110
helicobacter	pylori	infection	left	110
lung	cancer	mortality	left	110
protein	s	deficiency	left	110
prostate	cancer	cells	left	109
superoxide	dismutase	activity	left	109

Noun Compound			Bracketing	Freq.
b	cell	lines	left	108
blood	pressure	changes	left	108
contact	lens	wear	left	108
restriction	enzyme	analysis	left	108
somatostatin	receptor	scintigraphy	left	108
bone	marrow	aspirates	left	107
heparan	sulfate	proteoglycan	left	107
t	cell	hybridomas	left	107
collagen	type	iv	right	107
tumor	blood	flow	right	107
escherichia	coli	endotoxin	left	106
t	cell	receptor	left	106
tumor	cell	proliferation	left	106
color	doppler	ultrasound	right	106
protein	gene	product	right	106
carbon	monoxide	poisoning	left	105
heart	failure	patients	left	105
type	vii	collagen	left	105
intraclass	correlation	coefficients	right	105
mouse	bone	marrow	right	105
ribonuclease	protection	assay	left	105
attention	deficit	disorder	left	104
blood	pressure	responses	left	104
bronchoalveolar	lavage	fluid	left	104
growth	factor	stimulation	left	104
helicobacter	pylori	eradication	left	104
lipoprotein	lipase	activity	left	104
rat	liver	cytosol	left	104
signal	transduction	mechanisms	left	104
forearm	blood	flow	right	104
surface	plasmon	resonance	right	104
amino	acid	composition	left	103
bile	acid	synthesis	left	103
prostate	cancer	patients	left	103
ammonium	sulfate	fractionation	left	102
basement	membrane	components	left	102
body	weight	loss	left	102
mr	imaging	findings	left	102
nutrition	examination	survey	left	102
wall	motion	abnormalities	left	102
b	cell	proliferation	left	101
motor	vehicle	crashes	left	101

Noun Compound			Bracketing	Freq.
bone	marrow	samples	left	100
heart	transplant	patients	left	100
sister	chromatid	exchange	left	100
substance	abuse	treatment	left	100
type	iii	procollagen	left	100
rat	mast	cells	right	100
life	table	analysis	left	99
type	ii	receptors	left	99
flow	cytometry	analysis	left	98
gabaa	receptor	function	left	98
health	care	settings	left	98
protein	c	deficiency	left	98
stage	iii	disease	left	98
tnf	alpha	production	left	98
bone	marrow	examination	left	97
bone	marrow	transplants	left	96
gap	junction	channels	left	96
liver	biopsy	specimens	left	96
motor	neurone	disease	left	96
pressure	support	ventilation	left	96
tumour	suppressor	genes	left	96
tyrosine	kinase	receptors	left	96
escherichia	coli	lipopolysaccharide	left	95
house	dust	mites	left	95
inclusion	body	myositis	left	95
polymerase	chain	reactions	right	95
breast	cancer	screening	left	94
cell	surface	markers	left	94
growth	hormone	therapy	left	94
pertussis	toxin	treatment	left	94
x	chromosome	inactivation	left	94
dna	polymerase	delta	right	94
blood	glucose	control	left	93
bone	marrow	suppression	left	93
chicken	embryo	fibroblasts	left	93
group	b	streptococcus	left	93
health	system	reform	left	93
stage	ii	disease	left	93
blood	flow	velocities	left	92
guinea	pig	trachea	left	92
health	care	personnel	left	92
sprague	dawley	rats	left	92

Noun Compound			Bracketing	Freq.
transient	transfection	experiments	left	92
colon	carcinoma	cells	left	91
t	cell	receptors	left	91
blood	pressure	values	left	90
bone	marrow	biopsy	left	90
g	protein	activation	left	90
neutron	activation	analysis	left	90
portal	vein	thrombosis	left	90
protein	synthesis	inhibition	left	90
serine	protease	inhibitors	left	90
time	course	studies	left	90
varicella	zoster	virus	left	90
xanthine	oxidase	activity	left	90
bile	duct	injury	left	89
cat	scratch	disease	left	89
cell	surface	antigens	left	89
injection	drug	use	left	89
monoamine	oxidase	inhibitors	left	89
adenylyl	cyclase	activity	left	88
heart	rate	responses	left	88
type	x	collagen	left	88
urokinase	plasminogen	activator	left	88
family	planning	services	left	87
family	practice	residents	left	87
hepatitis	e	virus	left	87
messenger	rna	levels	left	87
organ	transplant	recipients	left	87
rna	polymerase	iii	left	86
ace	inhibitor	therapy	left	85
cell	cycle	kinetics	left	85
health	insurance	coverage	left	85
cancer	cell	lines	right	85
mouse	l	cells	right	85
amino	acid	transport	left	84
lv	ejection	fraction	right	84
aids	dementia	complex	left	83
cytochrome	oxidase	activity	left	83
potassium	channel	openers	left	83
injury	severity	score	left	82
iron	deficiency	anaemia	left	82
phospholipase	d	activity	left	82
t	cell	differentiation	left	82

Noun Compound			Bracketing	Freq.
borna	disease	virus	left	81
carbon	dioxide	production	left	81
dinucleotide	repeat	polymorphism	left	81
messenger	rna	expression	left	81
nadph	diaphorase	activity	left	81
nmda	receptor	function	left	81
rabbit	reticulocyte	lysate	left	81
t	lymphocyte	subsets	left	81
electrospray	mass	spectrometry	right	81
tissue	blood	flow	right	81
bone	marrow	aspiration	left	80
bone	marrow	failure	left	80
cobra	venom	factor	left	80
cyclobutane	pyrimidine	dimers	left	80
health	care	facilities	left	80
nerve	cell	bodies	left	80
cell	volume	regulation	left	79
ethidium	bromide	staining	left	79
superoxide	anion	generation	left	79
donor	t	cells	right	79
serum	hcv	rna	right	79
saccharomyces	cerevisiae	cells	left	78
host	defense	mechanisms	left	78
blood	pressure	elevation	left	77
emergency	room	visits	left	77
immunogold	electron	microscopy	right	77
breast	cancer	treatment	left	76
fibrin	degradation	products	left	76
glass	ionomer	cement	left	76
health	care	institutions	left	76
house	dust	mite	left	76
plasminogen	activator	activity	left	76
tyrosine	hydroxylase	activity	left	76
rat	sertoli	cells	right	76
cancer	pain	management	left	75
hiv	disease	progression	left	75
lactate	dehydrogenase	activity	left	75
t	cell	depletion	left	75
t	lymphocyte	activation	left	75
lh	pulse	frequency	right	75
polypeptide	growth	factors	right	75
inositol	phosphate	formation	left	74

Noun Compound			Bracketing	Freq.
phase	contrast	microscopy	left	74
tyrosine	kinase	activation	left	74
gc	b	cells	right	74
memory	b	cells	right	74
glutamate	receptor	antagonists	left	73
motor	nerve	terminals	left	73
t	cell	help	left	73
time	series	analysis	left	73
adult	wistar	rats	right	73
color	doppler	flow	right	73
protein	tyrosine	phosphatase	right	73
dna	strand	breakage	left	72
ground	reaction	forces	left	72
kidney	transplant	patients	left	72
integration	host	factor	right	72
murine	t	cells	right	72

Appendix D

Noun Phrase Coordination Dataset

Below I list the 428 test examples from the *Noun Phrase Coordination Dataset*, which were used in the experiments described in section 6.2.

Coordinated Noun Phrase				Coord. Type
cotton	and	acetate	fibers	noun
parts	and	marketing	operations	noun
sales	and	marketing	executive	noun
royalty	or	rock	stars	NP
Trade	and	Industry	Ministry	noun
Newsweek	and	U.S.	News	NP
chairman	and	chief	designer	NP
Securities	and	Exchange	Commission	noun
Taiwan	and	Saudi	Arabia	NP
investigations	and	trade	sanctions	NP
movie	and	book	pirates	noun
patent	and	copyright	owners	noun
Taiwan	and	Saudi	Arabia	NP
Washington	and	New	York	NP
Perch	and	Dolphin	fields	noun
Seahorse	and	Tarwhine	fields	noun
U.S.	and	Virgin	Islands	NP
capacity	and	debt	load	NP
president	and	chief	executive	NP
France	and	Hong	Kong	NP
U.S.	and	Soviet	Union	noun

Coordinated Noun Phrase				Coord. Type
milk	and	milk	powder	NP
practice	and	team	image	NP
business	and	computer	science	NP
analysts	and	business	people	NP
Taiwan	and	South	Korea	NP
business	and	research	program	noun
Public	and	Internatonal	Affairs	noun
Taiwan	and	South	Korea	NP
curriculum	and	testing	policies	NP
test	and	Learning	Materials	NP
test	and	Learning	Materials	NP
tests	and	practice	tests	NP
pie	and	bar	graphs	noun
tests	and	testing	format	NP
CAT	and	CTBS	tests	noun
insurance	and	services	concern	noun
Health	and	Human	Services	NP
Connecticut	and	Massachusetts	banks	noun
president	and	executive	officer	NP
review	and	advance	notice	NP
TV	and	movie	industry	noun
Health	and	Human	Services	NP
manufacturing	and	warehousing	space	noun
business	and	government	leaders	noun
research	and	development	facility	noun
polyproplene	and	polyester	film	NP
NL	and	Mr.	Simmons	NP
president	and	executive	officer	NP
NL	and	Mr.	Simmons	NP
NL	and	Mr.	Simmons	NP
government	and	business	leaders	noun
savings	and	investment	rates	noun
technology	and	market	knowledge	NP
activities	or	travel	clubs	NP
box	and	travelers	checks	NP
thrifts	and	credit	unions	NP
Securities	and	Exchange	Commission	noun
worth	and	employee	morale	NP
stock	and	futures	prices	NP
profits	and	cash	flow	noun
software	and	service	operations	noun
Mr.	and	Mrs.	Bush	noun

Coordinated Noun Phrase				Coord. Type
Triton	and	Mr.	Chase	NP
firm	and	Mr.	Whelen	NP
pulp	and	paper	segment	noun
credit	and	loan	guarantees	noun
cash	and	development	assistance	noun
Justice	and	Commerce	departments	noun
health	and	safety	violations	noun
maintenance	and	repair	programs	noun
safety	and	health	deficiencies	noun
Clairton	and	Fairless	works	noun
problems	and	substance	abuse	NP
cleanliness	and	health	care	NP
illness	and	substance	abuse	NP
Bricklayers	and	Allied	Craftsmen	NP
greed	or	profit	motive	NP
depressions	and	substance	abuse	NP
overproduction	and	distribution	problems	NP
power	and	appropriations	clause	NP
Pasadena	and	Long	Beach	NP
travel	and	entertainment	expenses	noun
Hammersmith	and	Fulham	council	noun
products	and	trading	techniques	noun
funds	and	pension	funds	NP
UAL	and	airline	shares	NP
stock	and	futures	markets	noun
futures	and	stock	markets	noun
buy	or	sell	order	noun
futures	or	program	trading	noun
FTC	and	Justice	Department	NP
FTC	and	Justice	Department	noun
Law	and	Democracy	Report	noun
environment	and	insurance	reform	NP
Stieglitz	and	Man	Ray	NP
horoscopes	and	romance	lines	NP
spot	and	futures	prices	noun
president	and	chief	executive	NP
terms	and	syndicate	manager	noun
Housing	and	Community	Development	noun
baseball	and	football	stadiums	noun
retail	or	customer	business	noun
miscarriages	and	morning	sickness	noun
Food	and	Drug	Administration	noun

Coordinated Noun Phrase				Coord. Type
California	and	New	York	NP
oil	and	mining	shares	noun
Securities	and	Exchange	Commission	noun
felony	and	misdemeanor	counts	noun
Philadelphia	and	Cleveland	districts	noun
estate	and	mortgage	loans	noun
Monopolies	and	Mergers	Commission	noun
stock	and	futures	markets	noun
buy	and	sell	orders	noun
disk	and	color	monitor	NP
March	and	May	contracts	noun
bond	and	stock	markets	noun
boots	and	leather	accessories	NP
Food	and	Drug	Administration	noun
television	and	radio	stations	noun
advertising	and	promotion	programs	noun
parts	and	aerospace	concern	noun
insurance	and	services	concern	noun
engineering	and	construction	company	noun
credit	and	loan	guarantees	noun
Israel	and	West	Germany	NP
construction	and	machinery	divisions	NP
Europe	and	Southeast	Asia	NP
president	and	chief	executive	noun
health	and	safety	violations	noun
chairman	and	chief	designer	noun
estate	and	equity	investments	NP
manufacturing	and	marketing	rights	noun
bone	or	cartilage	defects	noun
Taiwan	and	South	Korea	NP
France	and	West	Germany	NP
president	and	chief	executive	noun
steel	and	gas	operations	noun
takeover	or	business	combination	NP
steel	and	energy	segments	noun
president	and	executive	officer	NP
Securities	and	Exchange	Commission	noun
investors	and	money	managers	NP
stock	and	bond	fund	noun
futures	and	options	contracts	noun
repertoire	and	audience	appeal	noun
Mozart	and	Strauss	concertos	noun

Coordinated Noun Phrase				Coord. Type
servants	or	government	functionaries	NP
chemical	and	biotechnology	companies	noun
tennis	and	golf	tours	noun
accountants	and	estate	developers	NP
diesel	and	gas	turbines	noun
president	and	executive	officer	NP
economy	and	currency	issues	NP
president	and	executive	officer	NP
Treasury	and	Federal	Reserve	NP
mergers	and	acquisitions	department	noun
beer	and	dairy	products	NP
spirits	and	life	insurance	NP
Democrats	and	Bush	administration	NP
leaders	and	White	House	NP
House	and	Senate	leaders	noun
construction	and	operation	risks	noun
flooring	and	building	products	noun
military	and	security	agencies	NP
Digital	and	computer	makers	NP
Congress	and	Bush	administration	NP
Ways	and	Means	members	noun
instrument	and	controls	division	noun
insulation	and	fireproofing	concern	noun
rent	and	mortgage	payments	noun
rent	and	mortgage	payments	noun
rent	and	mortgage	payments	noun
rent	and	mortgage	payments	noun
rent	and	mortgage	payments	noun
steel	and	mining	community	noun
district	and	court	vacancies	noun
banks	and	bar	association	NP
cedar	and	brick	house	noun
Securities	and	Exchange	Commission	noun
terms	and	syndicate	manager	NP
chairman	and	executive	officer	NP
oil	and	gas	properties	noun
president	and	executive	officer	NP
business	and	development	philosophies	noun
managers	and	plant	workers	noun
compact	and	subcompact	cars	noun
criminals	and	street	people	NP
diapers	and	tissue	products	NP

Coordinated Noun Phrase				Coord. Type
pulp	and	newsprint	mill	noun
Oregon	and	South	Carolina	NP
Senate	and	House	bills	noun
treatment	and	tax	relief	NP
chairman	and	executive	officer	NP
Hyundai	and	J.P.	Morgan	NP
delver	and	detail	guy	NP
friends	and	tennis	partners	noun
premier	and	party	chief	noun
China	and	Soviet	Union	NP
Australia	and	New	Zealand	NP
Sihanouk	and	Hun	Sen	NP
brain	and	nerve	disorders	noun
London	and	New	York	NP
U.S.	and	West	Germany	NP
discount	and	Lombard	rates	noun
plunge	and	price	volatility	NP
asset	and	liability	management	noun
detective	and	security	agency	noun
conglomerate	or	investment	banker	NP
entertainment	and	publishing	businesses	noun
Temple	and	Sea	Containers	NP
points	and	closing	costs	NP
toiletries	and	plastic	bags	NP
finance	and	labor	committees	noun
consumer	and	provider	conceptions	noun
analysts	and	industry	consultants	NP
pulp	and	paper	prices	noun
chairman	and	executive	officer	NP
officials	and	utility	subsidiaries	NP
fat	and	cholesterol	content	noun
Tide	and	Mr.	Clean	NP
Japan	and	South	Korea	NP
Securities	and	Exchange	Commission	noun
pulp	and	paper	industry	noun
Food	and	Drug	Administration	noun
wage	and	benefit	costs	noun
banks	and	insurance	issues	noun
software	and	semiconductor	stocks	noun
bank	and	thrift	stocks	noun
Housing	and	Urban	Development	NP
Congress	and	White	House	NP

Coordinated Noun Phrase				Coord. Type
oil	and	gas	properties	noun
oil	and	gas	acquisitions	noun
oil	and	gas	revenue	noun
oil	and	gas	properties	noun
pipeline	and	oil	assets	noun
mining	and	energy	assets	noun
clothing	and	food	retailer	noun
tax	and	minority	interest	noun
steel	or	tin	plate	NP
marketing	and	cable	distribution	NP
refining	and	marketing	investment	noun
restructuring	or	merger	alternatives	noun
stocks	and	index	futures	NP
Food	and	Drug	Administration	noun
consulting	and	service	businesses	noun
Socialists	and	opposition	groups	NP
Immigration	and	Naturalization	Service	noun
companies	and	floor	traders	NP
dollar	and	bond	prices	NP
Securities	and	Exchange	Commission	noun
Lombard	and	discount	rates	noun
packaging	and	filtration	products	noun
gold	and	utility	shares	noun
inflation	and	stock	market	NP
gold	and	utility	issues	NP
refrigerator	and	freezer	manufacturer	noun
Industry	and	Government	Relations	noun
bricks	and	disk	drives	NP
banking	and	securities	businesses	noun
debt	and	equity	securities	NP
Securities	and	Exchange	Commission	noun
entertainment	and	news	shows	NP
chemicals	and	textiles	company	noun
Machinists	and	Aerospace	Workers	NP
chemicals	and	textile	businesses	noun
Congress	and	President	Bush	NP
MNC	and	Beneficial	offering	noun
research	and	development	costs	noun
Food	and	Drug	Administration	noun
allies	and	Soviet	Union	NP
technology	and	business	plans	noun
Securities	and	Exchange	Commission	noun

Coordinated Noun Phrase				Coord. Type
business	and	labor	leaders	noun
aid	and	food	grants	NP
Warner	and	Mr.	Azoff	NP
compensation	and	business	plans	NP
Ed	and	Lorraine	Warren	noun
crane	and	bulldozer	operators	noun
road	and	bridge	design	noun
marketing	or	computer	projects	noun
geology	and	petroleum	engineering	NP
wage	and	price	freeze	noun
hospital	and	insurance	company	noun
business	and	nursing	programs	noun
Securities	and	Exchange	Commission	noun
oil	and	gas	partnerships	noun
stock	and	bond	funds	noun
growth	and	growth	funds	NP
periodicals	and	community	newspapers	NP
beverages	and	consumer	products	NP
car	and	truck	production	noun
road	and	bridge	construction	noun
information	and	services	subsidiary	noun
U.S.	and	Western	Europe	NP
memory	and	processing	capability	NP
publishing	and	software	companies	noun
Rockefeller	and	Mellon	foundations	noun
aerospace	and	energy	concern	noun
Food	and	Drug	Administration	noun
U.S.	and	interest	rates	noun
Securities	and	Exchange	Commission	noun
banking	and	insurance	issues	noun
Bill	and	Bonnie	Quinlan	noun
lung	and	breast	cancers	noun
lung	and	breast	cancers	noun
brain	and	skin	cancers	noun
health	and	sanitation	concerns	noun
Minnesota	and	North	Dakota	NP
publishing	and	printing	concern	noun
energy	and	estate	company	noun
Food	and	Drug	Administration	noun
Boeing	or	work	force	NP
production	and	maintenance	workers	noun
oil	and	gas	producers	noun
Florida	and	North	Carolina	NP

Coordinated Noun Phrase				Coord. Type
terms	and	syndicate	manager	NP
Remics	and	mortgage	securities	NP
Paribas	and	Navigation	Mixte	NP
telecommunications	and	defense	equipment	noun
wire	and	cable	businesses	noun
oil	and	gas	acquisition	noun
exploration	and	production	unit	noun
reputation	and	sales	growth	noun
Securities	and	Exchange	Commission	noun
advertisers	and	advertising	agencies	NP
paper	and	plastic	products	noun
Visa	and	MasterCard	portfolio	noun
Visa	and	MasterCard	programs	noun
government	and	party	finances	noun
research	and	development	sectors	noun
banks	and	trading	companies	NP
developer	and	property	owner	noun
chairman	and	executive	officer	NP
losses	and	product	glitches	NP
chairman	and	chief	executive	NP
chairman	and	chief	executive	NP
bladder	and	urethra	problems	noun
wheat	and	beet	bran	NP
Food	and	Drug	Administration	noun
Kellogg	and	General	Mills	NP
investors	and	investment	banks	NP
stock	and	futures	markets	noun
stock	and	futures	markets	noun
bruises	and	cigarette	burns	NP
Recovery	and	Enforcement	Act	noun
professionals	and	community	leaders	NP
pipeline	and	marketing	concern	noun
communications	and	business	relationships	noun
relocation	and	severance	payments	noun
vaccine	and	bioresearch	firm	noun
Jimmy	and	Suburban	trucks	noun
Minpeco	and	Manufacturers	Hanover	NP
IRS	and	Manufacturers	Hanover	NP
IRS	and	Manufacturers	Hanover	NP
Securities	and	Exchange	Commission	noun
behemoths	and	robot	friend	NP
Netherlands	and	West	Germany	NP

Coordinated Noun Phrase				Coord. Type
polyester	and	rayon	markets	noun
policy	and	exchange	rates	NP
viability	and	ownership	changes	NP
volume	and	profit	impact	noun
food	and	tobacco	giant	noun
steam	and	combustion	turbines	noun
Friday	and	Saturday	nights	noun
scheduling	and	purchasing	personnel	noun
understanding	and	background	information	NP
data	and	cost	estimates	noun
songwriters	and	music	publishers	NP
Copyright	and	Home	Copying	noun
Defense	and	Space	Group	noun
Boeing	and	Machinists	representatives	NP
Aerospace	and	Electronics	groups	noun
Defense	and	Space	Group	noun
Aerospace	and	Electronics	division	noun
clients	and	media	people	NP
merger	or	tender	offer	noun
inflation	and	interest	rates	NP
chairman	and	chief	executive	NP
Hungary	and	Soviet	Union	NP
textile	and	clothing	section	noun
receivers	and	software	players	NP
steel	and	construction	shares	NP
Faberge	and	Elizabeth	Arden	NP
soap	and	toilet	paper	NP
research	and	development	bases	noun
lipstick	or	eye	makeup	NP
yardwork	and	home	maintenance	NP
insurers	or	Veterans	Administration	NP
Robert	and	Cynthia	Langendorf	noun
Securities	and	Exchange	Commission	noun
consumer	and	products	company	noun
color	and	food	additives	noun
directors	and	vice	presidents	NP
Florida	and	Puerto	Rico	NP
brokers	and	investment	bankers	NP
budgeting	and	planning	discipline	noun
U.S.	and	capital	markets	noun
terms	and	syndicate	manager	NP
trading	and	settlement	guidelines	noun
banking	and	currency	trading	NP

Coordinated Noun Phrase				Coord. Type
competition	and	client	needs	NP
Portugal	and	Puerto	Rico	NP
sales	and	dealer	orders	NP
tax	and	acccounting	practices	noun
wildlife	and	fishing	industry	NP
debt	and	debt	service	NP
production	and	pricing	differences	noun
hurricane	and	earthquake	damage	noun
Securities	and	Exchange	Board	noun
bonds	or	equity	shares	NP
petrochemical	and	agrochemical	company	noun
regulation	and	disclosure	requirements	noun
Securities	and	Exchange	Board	noun
media	and	resources	company	noun
Bond	and	Bell	Resources	noun
government	and	Jaguar	holders	NP
manufacturing	and	marketing	ventures	noun
banks	and	securities	companies	NP
beer	and	food	concern	noun
West	and	South	Carolina	NP
Austin	and	Fort	Worth	NP
Merger	and	acquisition	activity	noun
adviser	and	investment	banker	NP
film	and	television	operations	noun
cosmetology	and	business	schools	noun
Patterson	or	Sugarman	groups	noun
spending	and	capital	investment	NP

Appendix E

Comparing Human- and Web-Generated Paraphrasing Verbs

Below I list the Web- and the human-derived paraphrasing verbs for 250 noun-noun compounds from Levi (1978). The process of extraction is described in sections 4.4 and 4.7, respectively. For each noun-noun compound, I show the cosine correlation, the semantic class[1], and the paraphrasing verbs of each type, each followed by the corresponding frequency; the overlapping verbs are underlined. I show the results (top 10 verbs) by noun-noun compound when all human-proposed verbs are used and when only the first verb proposed by each worker is used in sections E.1 and E.2, respectively. I further show a comparison aggregated by semantic class (top 150 verbs) in sections E.3 and E.4, respectively.

[1] One of the recoverably deletable predicates proposed by Levi (1978), as described in section 2.5.1.

E.1 Comparison by Noun Compound: Using All Human-Proposed Verbs

0.96 "blood donor" NOMINALIZATION:AGENT

MTurk: give(30), donate(16), supply(8), provide(6), share(2), contribute(1), volunteer(1), offer(1), choose(1), hand over(1), ...

Web: give(653), donate(395), receive(74), sell(41), provide(39), supply(17), be(13), match(11), contribute(10), offer(9), ...

0.95 "women professors" BE

MTurk: be(22), teach(2), look like(2), be born(2), research(1), study(1), be comprised of(1), behave like(1), include(1), be gendered(1), ...

Web: be(251), teach(46), study(38), specialize in(26), appear as(13), think(10), take(10), research(9), work with(9), think that(9), ...

0.94 "student friends" BE

MTurk: be(18), come from(2), help(2), be had by(2), be made by(2), include(1), involve(1), act like(1), live as(1), work as(1), ...

Web: be(1250), have(34), support(33), know(31), teach(22), give(21), meet as(19), help(16), guide(15), pose as(12), ...

0.93 "city wall" $HAVE_2$

MTurk: surround(24), protect(10), enclose(8), encircle(7), encompass(3), be in(3), contain(2), snake around(1), border(1), go around(1), ...

Web: surround(708), encircle(203), protect(191), divide(176), enclose(72), separate(49), ring(41), be(34), encompass(25), defend(25), ...

0.91 "citizen soldier" BE

MTurk: be(20), represent(3), come from(3), act as(2), fight for(2), protect(1), defend(1), care for(1), start out as(1), guard(1), ...

Web: be(185), become(44), be treated as(14), save(14), dehumanize(11), shoot(10), rob(6), murder(5), fire upon(5), be drafted from(5), ...

0.91 "accident weather" $CAUSE_1$

MTurk: cause(21), promote(4), lead to(4), provoke(2), create(2), occur by(2), result in(2), contribute to(2), be condusive to(1), occasion(1), ...

Web: cause(25), provoke(3), be(3), include(2), end in(2), exist during(2), be proven with(1).

0.91 "disease germ" $CAUSE_1$

MTurk: cause(20), spread(5), carry(4), create(4), produce(3), generate(3), start(2), promote(2), lead to(2), result in(2), ...

Web: cause(919), produce(63), spread(37), carry(20), propagate(9), create(7), transmit(7), be(7), bring(5), give(4), ...

0.89 "flu virus" CAUSE_1

MTurk: cause(19), spread(4), give(4), result in(3), create(3), infect with(3), contain(3), be(2), carry(2), induce(1), ...

Web: cause(906), produce(21), give(20), differentiate(17), be(16), have(13), include(11), spread(7), mimic(7), trigger(6), ...

0.89 "collie dog" BE

MTurk: be(12), look like(8), resemble(2), come from(2), belong to(2), be related to(2), be called(2), be classified as(2), be made from(1), be named(1), ...

Web: be(24), look like(14), resemble(8), be border(5), feature(3), come from(2), tend(2), be bearded(1), include(1), betoken(1), ...

0.89 "gas stove" USE

MTurk: use(20), run on(9), burn(8), cook with(6), utilize(4), emit(3), be heated by(2), need(2), consume(2), work with(2), ...

Web: use(98), run on(36), burn(33), be(25), be heated by(10), work with(7), be used with(7), leak(6), need(6), consume(6), ...

0.87 "music box" MAKE_1

MTurk: play(19), make(12), produce(10), emit(5), create(4), contain(4), provide(2), generate(2), give off(2), include(1), ...

Web: play(104), make(34), produce(18), have(16), provide(14), be(13), contain(9), access(8), say(7), store(6), ...

0.87 "cooking utensils" FOR

MTurk: be used for(17), be used in(9), facilitate(4), help(3), aid(3), be required for(2), be used during(2), be found in(2), be utilized in(2), involve(2), ...

Web: be used for(43), be used in(11), make(6), be suited for(5), replace(3), be used during(2), facilitate(2), turn(2), keep(2), be for(1), ...

0.87 "honey bee" MAKE_1

MTurk: make(25), produce(16), create(9), manufacture(5), store(4), eat(2), provide(2), secrete(2), generate(1), gather(1), ...

Web: make(292), produce(189), gather(104), have(69), collect(57), suck(39), extract(19), drink(16), bring(15), carry(13), ...

0.85 "song bird" MAKE_1

MTurk: sing(22), chirp(4), warble(4), make(4), create(3), produce(3), be known for(2), generate(2), have(1), perform(1), ...

Web: sing(264), learn(80), have(69), hear(41), be(27), lose(18), continue in(18), know(15), develop(11), give(11), ...

0.85 "fruit tree" HAVE_1

MTurk: bear(20), produce(16), grow(15), have(6), give(4), provide(3), develop(2), supply(2), make(2), hold(1), ...

Web: bear(2113), produce(1215), have(337), bore(269), yield(201), bring(157), provide(155), bringeth(140), give(135), grow(74), ...

0.83 “concussion force” $CAUSE_1$

MTurk: cause(23), create(9), result in(5), produce(4), make(3), trigger(2), lead to(2), inflict(2), induce(2), form(2), ...

Web: cause(25), produce(15), make(2), suffer(2), be termed(1).

0.83 “servant girl” BE

MTurk: work as(17), be(14), act as(6), be employed as(5), serve as(3), be used as(2), be paid as(2), act like(2), be considered(2), have(1), ...

Web: be(193), work as(119), become(78), live as(20), require(16), live(10), be employed as(9), grow up as(9), go as(9), toil as(8), ...

0.82 “coke machine” FOR

MTurk: dispense(21), sell(19), contain(9), vend(7), distribute(6), hold(4), store(4), give(3), serve(2), supply(2), ...

Web: sell(25), dispense(16), give(9), offer(7), draw(7), supply(6), have(4), be camouflaged as(4), serve(3), carry(3), ...

0.80 “snow blindness” $CAUSE_2$

MTurk: be caused by(24), come from(7), result from(7), be created by(3), be derived from(2), emerge from(2), be induced by(2), be related to(2), come about from(1), be spawned by(1), ...

Web: be caused by(2), be like(1).

0.80 “desert rat” IN

MTurk: live in(22), reside in(10), be found in(6), come from(6), inhabit(5), survive in(3), exist in(3), thrive in(2), feed in(2), dwell in(2), ...

Web: live in(16), occur in(4), do(3), inhabit(2), love(1), survive in(1), inhibit(1).

0.79 “deficiency disease” $CAUSE_2$

MTurk: be caused by(18), come from(6), result from(5), be due to(3), be associated with(2), cause(2), start with(2), be created by(2), happen because of(2), stem from(1), ...

Web: be caused by(253), result from(203), be characterized by(81), have(69), cause(59), be associated with(56), involve(45), result in(40), be(39), arise from(34), ...

0.79 “mail sorter” NOMINALIZATION:AGENT

MTurk: sort(18), organize(9), handle(5), separate(4), work with(4), process(3), classify(3), divide(3), categorize(2), go through(2), ...

Web: sort(25), sweep(4), separate(1), deliver(1), result in(1).

0.78 “cigarette burn” $CAUSE_2$

MTurk: be caused by(21), come from(9), be made by(7), result from(6), be due to(4), originate from(2), be done with(2), emerge from(2), be left by(2), be created by(2), ...

Web: be caused by(6), come from(5), be from(4), consume(2), be inflicted by(1), be made by(1), resemble(1), suggest(1).

0.77 “novelty item” BE

MTurk: be(13), be called(4), be classified as(3), be considered(2), represent(2), have(2), resemble(2), be seen as(2), be sold as(2), inspire(2), ...

Web: be(14), include(4), display(3), range from(3), possess(2), appear as(2), be considered(1), be imroduced as(1), seem like(1).

0.77 “sap tree” MAKE$_1$

MTurk: produce(17), have(7), exude(5), contain(5), ooze(3), give(3), make(3), supply(2), create(2), secrete(2), ...

Web: produce(78), force(52), have(42), drip(33), exude(32), ooze(20), cause(18), get(17), secrete(16), yield(15), ...

0.77 “city folk” IN

MTurk: live in(21), reside in(13), come from(8), inhabit(5), stay in(2), be from(2), emerge from(2), be located in(2), be found in(2), occupy(2), ...

Web: live in(245), run(26), be in(22), leave(22), move from(22), come from(19), work in(17), flee(17), make(16), populate(13), ...

0.75 “farm boy” FROM

MTurk: work on(23), live on(22), come from(9), be born on(3), reside on(3), work in(3), grow up on(2), be raised on(2), toil on(2), labor on(2), ...

Web: live on(175), grow up on(103), work on(72), leave(67), remain on(13), be raised on(12), go from(11), be reared on(9), come from(8), live in(8), ...

0.75 “target structure” BE

MTurk: be(10), have(4), include(3), contain(3), involve(3), represent(2), resemble(2), provide(2), become(2), be used as(1), ...

Web: be(513), serve as(68), make(45), become(34), represent(21), provide(18), form(18), constitute(13), host(13), mimic(12), ...

0.73 “peer judgments” NOMINALIZATION:PRODUCT

MTurk: be made by(16), come from(11), be given by(4), involve(2), be rendered by(2), be from(2), originate from(1), be processed by(1), flow through(1), be handed down from(1), ...

Web: be made by(3), be had against(1).

0.73 “field mouse” IN

MTurk: live in(22), come from(9), be found in(7), inhabit(7), reside in(4), dwell in(3), breed in(2), habitates(2), hide in(2), burrow in(1), ...

Web: live in(38), ravage(10), be(8), suggest that(8), have(6), burrow in(5), infest(5), occur in(4), extend(4), be deprived of(3), ...

0.72 “beehive hairdo” BE

MTurk: look like(21), resemble(16), be shaped like(5), remind of(3), appear(2), emulate(2), be described as(1), be formed like(1), be referred to as(1), be modeled after(1), ...

Web: look like(4), be described as(4), resemble(2), be(1).

0.71 "vegetable soup" HAVE_1

MTurk: contain(14), be made from(11), have(9), be made of(8), include(5), taste like(4), use(3), consist of(3), be composed of(3), come from(3), ...

Web: contain(45), have(33), include(32), be made with(21), be(19), use(15), be filled with(10), be loaded with(8), puree(7), consist of(5), ...

0.70 "food supplies" NOMINALIZATION:PRODUCT

MTurk: include(12), contain(9), consist of(9), be(6), involve(4), be made of(4), be made up of(4), provide(3), be comprised of(3), have(2), ...

Web: include(58), be(34), make(17), consist of(11), provide(9), be used for(8), have(6), be derived from(6), mean that(5), flow around(5), ...

0.70 "sex scandal" ABOUT

MTurk: involve(15), be about(9), concern(4), revolve around(3), include(3), come from(3), be caused by(3), relate to(2), deal with(2), contain(2), ...

Web: involve(46), go beyond(21), revolve around(3), college(3), turn into(3), be(3), include(2), have(2), originate as(2), make(2), ...

0.69 "drug deaths" CAUSE_2

MTurk: be caused by(19), result from(7), be due to(6), occur because of(5), involve(2), come from(2), occur from(2), be related to(2), occur due to(2), be induced by(2), ...

Web: be caused by(25), be(14), involve(9), result from(9), stand for(8), be fueled by(7), mention(6), occur without(6), be blamed on(4), be associated with(4), ...

0.69 "college town" HAVE_1

MTurk: have(15), contain(15), house(4), surround(4), host(3), support(3), be near(2), depend on(2), center around(1), include(1), ...

Web: have(94), be(22), house(10), dissect(8), attend(7), host(6), contain(5), support(5), center around(4), accredit(4), ...

0.68 "grain alcohol" FROM

MTurk: be made from(16), come from(10), be produced from(7), be distilled from(5), contain(5), use(3), originate from(2), be made of(2), be fermented from(2), be derived from(1), ...

Web: be made from(13), be produced from(11), be manufactured from(6), be derived from(5), require(4), crystallize in(3), raise(3), be called(1), mean(1), produce from(1), ...

0.68 "flounder fish" BE

MTurk: be(9), look like(6), resemble(3), be called(3), be classified as(3), taste like(2), be identified as(2), be named(1), smell like(1), group with(1), ...

Web: look like(9), be(4), resemble(4), include(2), be used include(2), substitute for(1), go with(1), be shaped like(1), follow(1), want(1).

0.68 "growth hormone" CAUSE_1

MTurk: promote(11), cause(9), stimulate(6), affect(4), encourage(3), help(3), increase(2), regulate(2), enhance(2), produce(2), ...

Web: stimulate(403), regulate(222), promote(211), affect(105), control(67), cause(61), inhibit(31), influence(28), include(28), encourage(22), ...

0.67 "government land" HAVE_2

MTurk: be owned by(23), belong to(15), be managed by(4), be controlled by(3), be possessed by(3), be administered by(2), come from(2), be maintained by(2), be occupied by(1), be regulated by(1), ...

Web: be owned by(205), be acquired by(80), have(58), be(41), be taken by(39), be purchased by(34), be granted by(27), be leased from(26), be held by(23), be acquired from(21), ...

0.67 "picture book" HAVE_1

MTurk: contain(19), have(10), include(6), use(4), be composed of(3), be made up of(3), display(3), consist of(3), utilize(3), hold(2), ...

Web: have(1146), contain(366), include(345), give(290), show(201), be(157), paint(119), present(68), use(67), provide(60), ...

0.66 "film cutter" NOMINALIZATION:AGENT

MTurk: cut(16), edit(11), slice(9), divide(3), trim(3), work with(3), snip(2), splice(2), chop up(1), separate(1), ...

Web: cut(20), see(7), divide(5), chop up(3), conform(3), separate(2), sever(2), perforate(2), act in(1), be reciprocated across(1), ...

0.64 "pet families" HAVE_1

MTurk: have(14), own(13), love(8), care for(6), consist of(5), like(4), enjoy(4), be composed of(4), contain(4), include(2), ...

Web: have(187), adopt(84), lose(77), love(46), want(43), own(38), bring(35), keep(28), include(24), take(24), ...

0.64 "picture album" FOR

MTurk: contain(22), hold(16), display(11), have(5), protect(4), show(3), store(3), organize(3), be filled with(2), show off(2), ...

Web: contain(318), have(310), include(193), need(42), be(36), hold(34), feature(31), show(29), be filled with(28), display(19), ...

0.63 "hand brake" USE

MTurk: be operated by(11), be used by(7), require(6), use(5), be activated by(5), be made for(4), be applied by(3), operate by(3), be controlled by(3), need(3), ...

Web: be operated by(112), be applied by(14), be controlled by(6), be(6), be set by(5), be released by(5), require that(4), require(3), be of(3), secure(3), ...

0.63 "cactus plant" BE

MTurk: be(12), look like(9), be called(2), grow like(2), be related to(2), contain(2), be classified as(2), resemble(1), originate from(1), be composed of(1), ...

Web: resemble(83), look like(77), be(38), include(13), grow like(7), be confused with(5), last(4), be among(4), associate with(3), range from(3), ...

0.62 "shock treatment" USE

MTurk: use(16), utilize(7), involve(6), consist of(6), cause(3), include(2), induce(2), give(2), create(2), be comprised of(2), ...

Web: include(13), use(13), involve(8), be(8), cause(5), produce(4), remove(4), be used during(3), induce(2), ensure(2), ...

0.62 “head noun” BE

MTurk: be(6), describe(3), involve(2), be found in(2), refer to(2), be at(2), come at(2), stand at(1), have(1), form(1), . . .

Web: be(149), form(24), function as(16), occur as(14), appear as(13), modify(10), act as(7), serve as(7), precede(5), be analysed as(4), . . .

0.62 “designer creations” NOMINALIZATION:PATIENT

MTurk: be made by(20), come from(6), be designed by(6), be produced by(5), be created by(4), be presented by(2), be from(2), originate from(2), be developed by(2), be thought of by(2), . . .

Web: be made by(5), be produced by(4), be selected with(4), be created by(3), leave(3), require(3), come from(2), be from(2), land(2), design(2), . . .

0.62 “tear gas” CAUSE_1

MTurk: cause(18), create(9), make(7), generate(5), produce(3), induce(3), provoke(2), initiate(1), lead to(1), encourage(1), . . .

Web: cause(13), choke as(8), be like(6), include(4), tamponades(4), be(4), induce(3), be made from(3), trigger(2), bring(2), . . .

0.61 “country butter” FROM

MTurk: be made in(18), come from(14), be produced in(6), be associated with(3), be from(3), originate in(2), be churned in(2), taste(1), be styled after(1), stand for(1), . . .

Web: be produced in(12), come from(8), be made in(3), be preferred in(3), be imported into(3), be picked up around(2), be consumed in(2), be celebrated around(2), enter(1), leave(1), . . .

0.60 “automobile plant” FOR

MTurk: make(17), assemble(15), manufacture(15), build(11), produce(10), create(3), turn out(2), put together(2), construct(2), manufatures(1), . . .

Web: produce(16), be utilized in(14), build(7), manufacture(7), assemble(5), include(5), make(3), turn out(2), be used in(2), process(2), . . .

0.59 “winter season” BE

MTurk: be(7), include(6), occur in(5), be in(5), occur during(4), feel like(2), happen during(2), be called(2), come in(2), exist in(2), . . .

Web: be(128), occur in(20), be during(17), be in(15), follow(15), coincide with(14), be called(13), include(12), join(10), run from(9), . . .

0.59 “rye whiskey” FROM

MTurk: be made from(18), come from(7), contain(5), be made of(5), be brewed from(4), be distilled from(3), use(3), be derived from(2), be formulated from(2), be created from(2), . . .

Web: be distilled from(4), be made from(3), contain(1), call(1), accent(1), have(1).

0.58 “family problems” IN

MTurk: involve(5), affect(5), occur within(3), plague(3), happen in(3), be experienced by(2), occur in(2), be found in(2), be solved by(2), arise in(2), . . .

Web: affect(318), plague(120), arise in(90), bring(90), confront(79), occur in(51), run in(47), involve(41), cause(39), threaten(33), . . .

0.56 "sports activities" BE

MTurk: involve(12), include(7), be(4), be related to(4), use(3), revolve around(2), be comprised of(2), contain(2), play(1), focus on(1), . . .

Web: include(494), be(118), range from(70), involve(62), use(29), be considered(29), bring(16), organize(13), be classified as(13), be associated with(12), . . .

0.55 "worker teams" MAKE_2

MTurk: be made up of(11), include(10), be composed of(6), be comprised of(6), have(5), consist of(5), contain(4), comprise(3), employ(3), be made from(3), . . .

Web: include(407), consist of(94), be(31), have(13), bring(10), provide(10), give(8), allow(8), incorporate(7), need(7), . . .

0.54 "mining engineer" FOR

MTurk: work in(9), study(9), specialize in(6), know about(6), do(3), work with(3), design(3), understand(2), know(2), investigate(2), . . .

Web: specialize in(15), work in(10), be(7), stay within(6), work for(6), do(5), work with(5), build(3), select(3), understand(2), . . .

0.54 "bear country" HAVE_1

MTurk: have(11), contain(9), be inhabited by(6), house(4), harbor(4), support(4), feed(3), protect(2), supply(2), sustain(2), . . .

Web: have(33), do(15), protect(10), be(6), go(5), sustain(3), be screened in(3), deliver(3), see as(3), allow(3), . . .

0.54 "star shape" BE

MTurk: resemble(18), look like(16), be like(3), appear like(3), emulate(2), describe(1), depict(1), represent(1), form(1), be made from(1), . . .

Web: include(16), resemble(11), look like(8), be(6), represent(5), seep down from(5), form(4), blot out(4), create as(4), vary from(4), . . .

0.54 "lightning rod" FOR

MTurk: attract(19), protect from(7), conduct(6), channel(5), draw(3), deflect(3), direct(3), divert(2), absorb(2), dissipate(2), . . .

Web: attract(17), be struck by(13), be melted by(9), convey(9), turn away(8), prevent(5), conduct(3), capture(3), ground(2), bear(2), . . .

0.53 "lion cub" BE

MTurk: be(9), come from(9), grow into(6), become(5), be born of(4), be birthed by(3), be born from(3), bear of(2), belong to(2), resemble(2), . . .

Web: become(6), be(4), grow into(3), blurt out(2), meet with(2), follow(2), belong in(2), have(2), make up(1), be threatened by(1), . . .

0.53 "faculty decisions" NOMINALIZATION:PRODUCT

MTurk: be made by(25), come from(5), affect(3), involve(2), be decided by(2), concern(2), include(1), be reserved for(1), be made in(1), be engendered by(1), . . .

Web: affect(60), be made by(33), go against(9), maintain(7), allow(7), be criticized by(6), be(6), be supported by(5), restrict(5), rest with(5), . . .

0.53 "oil well" FOR

MTurk: contain(15), produce(10), pump(5), provide(4), have(3), draw(3), hold(2), gush(2), spew(2), release(2), ...

Web: produce(689), find(78), test(58), pump(44), be(35), flow(34), be drilled for(32), contain(23), yield(22), have(21), ...

0.52 "cane sugar" FROM

MTurk: come from(18), be made from(11), be derived from(5), be produced from(5), be created from(4), be processed from(4), be extracted from(3), be made of(3), originate from(2), be squeezed from(2), ...

Web: be(18), come from(14), be made from(6), replace(6), be refined(5), predominate over(5), be in(4), be unlike(4), be derived from(3), compete with(3), ...

0.52 "child actor" BE

MTurk: be(22), look like(4), portray(3), start as(1), include(1), play(1), have(1), involve(1), act like(1), star as(1), ...

Web: play(113), be(65), have(32), portray(31), start as(18), work with(18), deal with(17), write(17), look like(15), make(13), ...

0.51 "fish scales" $HAVE_2$

MTurk: cover(11), protect(8), be on(7), weigh(6), be found on(6), grow on(6), come from(5), appear on(2), coat(2), shed from(1), ...

Web: be found on(17), be embedded in(17), cover(14), weigh(12), sustain(9), make(7), attract(7), look like(5), allow(5), protect(4), ...

0.50 "nut bread" $HAVE_1$

MTurk: contain(21), include(10), be made with(9), have(8), be made from(5), use(3), be made using(3), feature(2), be filled with(2), taste like(2), ...

Web: have(8), include(5), feature(4), be mixed with(4), be crammed with(3), load with(3), contain(2), be filled with(2), be stuffed with(2), be loaded with(2), ...

0.50 "neighborhood bars" IN

MTurk: be in(10), be located in(10), be found in(4), be situated in(3), serve(2), protect(2), be populated by(2), be used by(2), reside in(2), exist in(2), ...

Web: be in(12), open in(8), ground in(5), typify(5), be located in(4), rock(4), pop up in(3), open up in(3), shutter in(3), give(3), ...

0.49 "sports magazine" ABOUT

MTurk: discuss(9), cover(6), feature(5), be about(5), write about(4), talk about(4), tell about(3), report on(2), involve(2), look at(2), ...

Web: cover(95), feature(46), focus on(26), include(13), promote(12), deal with(11), depict(9), make(8), be(8), have(5), ...

0.49 "country visitors" FROM

MTurk: go to(9), come from(8), visit(8), travel to(7), live in(3), tour(3), hail from(3), be from(3), enjoy(2), explore(2), ...

Web: come from(135), enter(133), visit(44), be in(42), leave(36), be from(26), come into(24), reside in(21), live in(18), stay in(18), ...

0.48 "child abuse" NOMINALIZATION:ACT

MTurk: harm(9), affect(9), happen to(8), involve(7), hurt(6), be done to(5), concern(4), impact(3), occur to(3), afflict(2), ...

Web: affect(76), cause(30), place(30), be(29), involve(27), put(27), harm(26), drive(25), teach(23), be inflicted upon(20), ...

0.48 "fatigue headache" CAUSE$_2$

MTurk: be caused by(19), result from(9), come from(7), be due to(4), be brought on by(4), be produced by(3), happen because of(3), indicate(2), be related to(2), follow(2), ...

Web: be induced by(3), be caused by(2), result from(2), result in(2), be preceded by(2), arise from(1), be aggravated by(1).

0.48 "winter sports" IN

MTurk: be played in(17), occur in(9), happen in(7), occur during(5), be played during(4), happen during(3), require(3), be done in(2), celebrate(2), be in(2), ...

Web: be played in(7), move from(6), be done in(4), enjoy in(3), be prepared by(3), be played during(2), be divided into(2), come with(2), dominate(2), play(2), ...

0.47 "sister node" BE

MTurk: be(4), have(3), be like(3), act like(2), be described as(2), act as(2), be related like(2), accompany(2), belong to(1), hook up with(1), ...

Web: be(105), have(4), be contained in(4), be inserted as(3), be adjoined as(3).

0.47 "pine tree" BE

MTurk: be(6), have(5), smell like(4), be called(4), produce(4), smell of(3), be made of(3), be made from(2), grow(2), be classified as(2), ...

Web: be(88), include(37), resemble(25), look like(18), surpass(10), be called(6), smell like(6), be planted(6), replace(6), crowd(6), ...

0.47 "garter snake" BE

MTurk: look like(13), resemble(9), be called(2), have(2), appear(2), be named after(2), eat(1), like(1), take away(1), be thin like(1), ...

Web: resemble(5), be(5), include(4), look like(2), find(2), be triggered by(2), be confused with(2), eat(1), be construed as(1), sell(1).

0.46 "finger cymbals" USE

MTurk: be played with(8), be worn on(8), be played by(5), be used by(4), use(3), attach to(3), go on(2), resemble(2), need(2), be placed on(2), ...

Web: go on(1), be played with(1).

0.45 "marine life" IN

MTurk: be(9), exist in(4), live in(3), be enjoyed by(2), dwell in(2), originate from(2), be lived by(2), be of(2), be found in(2), be described as(2), ...

Web: be(3), help(3), use(2), support(1), exist in(1), characterize(1), float in(1), showcases(1), end(1).

0.45 “coal dust” FROM

MTurk: come from(20), be made from(5), be(4), be made of(4), be created by(4), derive from(4), be generated from(2), be composed of(2), emit from(2), be produced by(2), ...

Web: come from(5), be associated with(5), fuel(5), be generated from(4), contain(4), be(4), fall from(3), be caused from(2), feed(2), be correlated with(2), ...

0.45 “plastic toys” MAKE_2

MTurk: be made of(18), contain(11), be made from(7), be composed of(6), consist of(4), be manufactured from(3), come from(3), be constructed of(2), use(2), be built of(2), ...

Web: be(51), be made of(25), have(10), be made from(8), consist of(4), be made out of(3), heat(3), fire(3), be constructed of(2), be constructed from(2), ...

0.45 “city planner” NOMINALIZATION:AGENT

MTurk: design(11), plan(10), work for(5), organize(4), build(3), envision(3), lay out(3), work with(3), be concerned with(3), develop(2), ...

Web: work with(27), work for(24), serve on(22), work in(14), design(13), treat(12), build(11), work at(9), understand(7), deal with(7), ...

0.44 “college employees” NOMINALIZATION:PATIENT

MTurk: work for(15), work at(13), be employed by(11), work in(5), be paid by(5), be hired by(4), run(2), be in(2), be used by(2), go to(2), ...

Web: leave(95), be employed by(69), graduate from(63), attend(50), work for(32), serve(29), work at(26), have(23), be enrolled in(22), retire from(22), ...

0.42 “frog man” BE

MTurk: resemble(11), swim like(9), look like(6), act like(4), be like(3), be(2), dive like(2), emerge from(2), sell(1), study(1), ...

Web: look like(24), be(12), find(12), resemble(11), hate(7), live on(7), blow up(7), swallow(7), find frog(7), write about(5), ...

0.42 “chocolate bar” MAKE_2

MTurk: contain(17), be made of(16), be made from(10), taste like(7), be composed of(7), consist of(5), be(3), have(2), smell of(2), be manufactured from(2), ...

Web: be(103), contain(54), have(30), serve(26), taste like(13), combine(11), be made with(7), be sold in(6), be called(6), come in(6), ...

0.41 “suspense film” CAUSE_1

MTurk: contain(9), have(7), create(6), cause(6), include(3), generate(3), involve(3), be(2), utilize(2), produce(2), ...

Web: rely on(26), have(24), build(22), create(16), combine(14), use(9), be(8), lack(8), miss(8), offer(7), ...

0.40 “census taker” NOMINALIZATION:AGENT

MTurk: take(7), count(4), record(3), work for(3), collect(3), administer(2), work on(2), compile(2), poll for(2), conduct(2), ...

Web: take(2), mean that(2).

0.40 "sob story" CAUSE_1

MTurk: cause(12), make(8), create(7), induce(6), lead to(6), produce(5), encourage(4), elicit(4), evoke(4), generate(3), ...

Web: make(15), produce(2), be interrupted by(1), turn into(1), come through(1).

0.39 "cable network" MAKE_2

MTurk: consist of(5), be made of(5), be on(4), use(4), provide(3), run on(3), be linked by(2), carry(2), be watched on(2), be made up of(2), ...

Web: use(154), include(52), be(31), consist of(25), provide(19), run over(18), employ(15), run(15), have(13), compete in(13), ...

0.36 "peanut butter" FROM

MTurk: be made from(18), come from(9), contain(7), taste like(6), be made of(5), be produced from(3), consist of(3), be derived from(2), include(2), use(2), ...

Web: contain(34), be(26), have(19), be made from(7), taste like(5), use(4), be than(4), resemble(3), list(3), produce(3), ...

0.35 "oil imports" NOMINALIZATION:PRODUCT

MTurk: be(11), contain(8), involve(6), include(4), supply(4), bring in(4), be made from(3), consist of(2), be made of(2), comprise(2), ...

Web: include(23), consist of(5), be(5), constitute(5), result in(4), be of(3), be purchased through(3), exclude(3), offset(2), account for(2), ...

0.35 "cream sauce" HAVE_1

MTurk: contain(11), be made from(8), use(6), be made of(6), include(4), consist of(3), be composed of(3), involve(2), be like(2), resemble(2), ...

Web: combine(19), have(17), be(14), include(8), contain(6), call for(6), taste like(5), be enriched with(5), need(5), be made from(4), ...

0.35 "home remedy" FROM

MTurk: be made at(12), come from(11), be used at(6), be used in(4), be found at(3), originate at(3), originate from(2), happen at(2), be developed at(2), be prepared at(1), ...

Web: be prepared at(16), be in(14), be tried at(11), be made at(10), be used in(6), be done at(6), be administered at(4), be used at(4), build(4), be(4), ...

0.35 "love song" ABOUT

MTurk: be about(17), describe(6), talk about(5), concern(4), involve(3), speak of(3), refer to(3), demonstrate(2), deal with(2), mention(2), ...

Web: make(127), be(123), do(59), be about(58), celebrate(49), express(48), deal with(40), describe(33), speak of(32), go(32), ...

0.35 "student committee" MAKE_2

MTurk: be made up of(10), be composed of(8), be comprised of(7), involve(6), contain(6), consist of(6), be made of(5), include(4), be run by(4), empower(3), ...

Web: include(523), consist of(123), allow(100), be comprised of(76), meet with(65), select(64), be(64), advise(62), guide(60), be composed of(57), ...

0.34 “sea breeze” FROM

MTurk: come from(22), blow from(4), smell like(4), be near(3), blow off(2), originate from(2), come off of(2), emerge from(2), be caused by(2), rise from(1), . . .

Web: blow from(72), come from(16), come off(15), blow towards(13), ruffle(13), make(12), cause(12), occur in(9), frilled(9), come over(8), . . .

0.33 “paper money” MAKE$_2$

MTurk: be made of(16), be made from(13), be printed on(6), be(5), be manufactured from(5), consist of(4), be composed of(4), resemble(3), be comprised of(3), feel like(2), . . .

Web: be(22), exist on(12), be spent on(11), be made from(6), exist in(5), give(5), be wrapped in(5), remain(5), pass along(5), say(5), . . .

0.33 “teaching profession” BE

MTurk: involve(12), include(7), require(7), be about(4), promote(2), focus on(2), revolve around(2), center on(2), use(2), espouse(2), . . .

Web: be(52), include(23), involve(12), have(12), be engaged in(10), enjoy(6), require(5), be liberalized by(5), recognize(5), promote(3), . . .

0.33 “queen bee” BE

MTurk: be(13), act as(3), be called(2), rule as(2), serve as(2), reign as(2), act like(2), be positioned as(1), be relegated(1), develop into(1), . . .

Web: have(26), lose(26), be(19), surround(9), serve(9), feed(8), become(8), be with(7), be without(7), be separated from(7), . . .

0.33 “store clothes” FROM

MTurk: come from(11), be sold in(10), be bought in(8), be found in(5), be purchased at(5), be worn in(4), be purchased in(4), be bought from(3), be purchased from(3), be sold by(2), . . .

Web: be in(46), be sold in(29), sell in(11), arrive in(10), come into(9), hit(9), be sold by(6), be sold at(6), go into(5), fit in(5), . . .

0.33 “extension ladder” HAVE$_1$

MTurk: have(16), provide(5), use(4), contain(4), include(3), supply(3), make(3), form(2), perform(2), utilize(2), . . .

Web: be(12), be used as(8), have(6), form(4), be generated by(4), be designed as(3), double as(3), be arranged in(3), be like(2), be utilized as(2), . . .

0.33 “bacon grease” FROM

MTurk: come from(22), be derived from(5), be rendered from(4), be made from(4), taste like(3), be produced by(3), be from(3), smell like(2), be contained in(2), be found in(2), . . .

Web: be rendered from(4), run off(2), come from(1), sit on(1).

0.32 “adventure story” ABOUT

MTurk: contain(8), be about(7), involve(6), describe(6), include(4), tell of(4), have(4), relate(3), tell about(3), discuss(3), . . .

Web: follow(124), be(54), tell of(51), recount(49), relate(49), combine(37), chronicle(34), continue(31), tell(29), describe(27), . . .

0.31 "mountain lodge" IN

MTurk: be in(9), be located in(9), be built in(6), be on(5), be found in(5), be situated on(4), be built on(2), be found on(2), be near(2), be nestled in(1), . . .

Web: have(6), be located in(4), curse(4), be in(3), be nestled in(3), sit atop(3), operate on(3), overlook(3), be styled like(3), be tucked in(2), . . .

0.31 "candy cigarette" $MAKE_2$

MTurk: be made of(15), taste like(8), be made from(7), be(6), look like(5), be composed of(3), resemble(3), contain(3), create from(2), be eaten like(2), . . .

Web: taste like(3), have(2).

0.31 "steam iron" USE

MTurk: use(12), produce(7), emit(6), create(5), make(4), utilize(4), work by(3), function with(2), need(2), give off(2), . . .

Web: generate(14), deliver(7), emit(4), produce(4), do(4), reduce(4), spray(3), discharge(3), use(3), operate on(3), . . .

0.30 "future shock" $CAUSE_2$

MTurk: occur in(6), involve(5), happen in(5), come from(4), be caused by(4), concern(3), be about(3), happen during(2), happen because of(2), result from(2), . . .

Web: occur in(10), affect(8), persist in(5), reverberate into(3), be(3), present(2), dissipate in(2), come in(1), be produced in(1), determine(1), . . .

0.30 "copper coins" $MAKE_2$

MTurk: be made of(12), be made from(10), contain(9), be composed of(5), consist of(4), be cast from(3), be made out of(3), include(2), be produced from(2), be comprised of(2), . . .

Web: be(36), be made of(14), have(7), be of(7), contain(4), be made from(3), be struck in(3), look like(3), depress(3), show(3), . . .

0.30 "tuition subsidies" NOMINALIZATION:PRODUCT

MTurk: pay for(8), fund(5), subsidize(4), cover(4), help with(4), support(4), pay(3), lower(3), provide(3), provide for(2), . . .

Web: keep(7), pay for(4), enable(3), lower(2), approximate(2), include(2), pay(1), reduce(1), make(1).

0.30 "policy matters" ABOUT

MTurk: involve(7), concern(7), be related to(4), relate to(4), pertain to(4), refer to(3), deal with(3), be about(3), regard(2), discuss(2), . . .

Web: affect(134), involve(75), be(52), be covered by(44), have(33), include(17), depend upon(14), require(13), write(12), conflict with(11), . . .

0.30 "satellite nation" BE

MTurk: act like(5), be(4), behave like(3), consist of(3), have(2), use(2), be like(2), resemble(2), be considered(2), be governed as(1), . . .

Web: launch(23), be(20), put(7), use(6), have(5), exist(5), bring(5), be transmitted via(4), depend on(4), be governed as(3), . . .

0.29 "city trainees" NOMINALIZATION:PATIENT

MTurk: work for(13), be employed by(7), be trained by(7), work in(6), live in(4), learn from(3), be trained in(3), be from(3), work at(2), be trained for(2), ...

Web: work in(6), live outside(4), work at(3), complete(3), be hired by(2), be(2), be employed by(1), join(1).

0.29 "headache pills" FOR

MTurk: cure(14), relieve(8), reduce(5), help(4), treat(4), ease(3), do away with(2), decrease(2), alleviate(2), heal(2), ...

Web: give(38), cause(19), cure(15), take away(9), develop(8), stop(5), relieve(5), rid of(4), prevent(3), be for(3), ...

0.28 "bronze statue" MAKE$_2$

MTurk: be made of(20), be composed of(5), be(5), contain(5), be cast from(4), be made from(3), be cast of(2), be manufactured from(2), be created with(2), be sculpted in(2), ...

Web: be cast in(52), be of(29), be(20), be made of(12), be in(9), be cast from(6), be cast of(3), become(3), inform(3), penetrate(3), ...

0.28 "apple cake" HAVE$_1$

MTurk: contain(16), be made from(10), be made with(9), be made of(7), have(6), taste like(6), include(4), use(3), come from(3), taste of(2), ...

Web: use(2), have(2), substitute(2), consist of(1), include(1), be made with(1), chop(1), dry(1), be(1), be shaped like(1), ...

0.28 "enemy invasion" NOMINALIZATION:ACT

MTurk: involve(10), come from(8), be made by(5), be done by(3), be started by(3), be caused by(3), be initiated by(2), have(2), be accomplished by(2), emerge from(2), ...

Web: be made by(4), make(4), be kept upbj(4), come from(3), be attempted by(3), drive(3), control(2), be aimed at(2), be visited upon(2), be held out by(1), ...

0.28 "company assets" HAVE$_2$

MTurk: belong to(16), be owned by(14), be held by(3), be used by(3), be generated by(2), lie within(2), be possessed by(2), reside in(1), involve(1), be produced by(1), ...

Web: enable(173), allow(104), be owned by(77), require(62), affect(60), be used by(56), be adopted by(55), be held by(51), be(48), increase(39), ...

0.28 "draft dodger" NOMINALIZATION:AGENT

MTurk: avoid(22), run from(9), evade(7), hide from(5), escape(5), run away from(4), elude(4), duck(2), object to(2), ignore(2), ...

Web: dodge(7), evade(6), defy(2), protest(2), avoid(1), prosecute(1).

0.27 "cover designs" NOMINALIZATION:PRODUCT

MTurk: be on(7), decorate(7), appear on(5), be made for(5), grace(2), create(2), adorn(2), be used for(2), be found in(2), be created for(2), ...

Web: allow(70), appear on(42), grace(33), have(16), become(16), make(15), feature(13), include(12), do(10), be used for(9), ...

0.26 "price dispute" ABOUT

MTurk: be about(12), involve(10), concern(8), be over(7), question(3), regard(2), argue(2), center on(2), stem from(2), argue about(2), ...

Web: regard(5), respect(3), translate into(3), be about(2), challenge(2), break out over(2), push(2), increase(1), concern(1), exist over(1), ...

0.26 "cell division" NOMINALIZATION:ACT

MTurk: occur in(8), split(7), divide(6), happen to(5), multiply(4), happen in(4), reproduce(4), involve(3), affect(3), result in(2), ...

Web: produce(173), result in(60), generate(51), occur in(38), divide(36), form(17), cause(14), turn(12), make(11), create(11), ...

0.25 "vacuum cleaner" USE

MTurk: use(12), clean by(6), create(3), utilize(3), have(2), need(2), suck into(2), work with(2), apply(2), require(2), ...

Web: be(44), invent(12), use(11), be like(7), make(6), produce(6), have(5), resemble(5), pick up than(5), include(4), ...

0.25 "mother church" BE

MTurk: act as(8), act like(6), be(3), be like(3), serve as(2), feel like(2), resemble(2), behave as(2), emulate(2), be supported by(1), ...

Web: be(447), support(20), become(19), have(16), be brought up with(16), be represented as(11), help(10), be like(9), tell(8), find(7), ...

0.25 "student decision" NOMINALIZATION:ACT

MTurk: be made by(22), come from(6), affect(5), involve(4), originate from(3), be made about(2), belong to(2), include(2), be taken by(2), be about(2), ...

Web: affect(634), allow(80), support(42), benefit(41), be made by(31), require(29), involve(21), enable(21), impact(21), limit(18), ...

0.25 "mountain range" MAKE$_2$

MTurk: be made of(12), contain(8), be composed of(5), have(5), include(3), be made up of(3), be of(2), comprise(2), consist of(2), be comprised of(2), ...

Web: be(61), include(41), connect(16), have(15), contain(8), build with(8), be among(8), make(6), comprise(5), extend from(5), ...

0.24 "juvenile court" FOR

MTurk: try(13), deal with(8), judge(5), handle(4), adjudicate(3), prosecute(3), process(3), be for(3), involve(2), sentence(2), ...

Web: include(29), state that(29), be(25), place(18), commit(18), deal with(15), adjudicate(13), handle(13), sit as(12), confirm(9), ...

0.24 "birth pains" CAUSE$_2$

MTurk: be caused by(11), occur during(6), accompany(5), result from(4), be due to(4), precede(2), come from(2), coincide with(2), occur because of(2), develop during(2), ...

Web: accompany(55), give(55), precede(16), follow(16), attend(14), come with(11), be suspended after(10), occur from(7), herald(5), arise from(4), ...

0.24 “childhood dreams” IN

MTurk: occur in(6), occur during(6), happen during(5), happen in(4), come from(3), originate in(3), come in(3), start in(2), relate to(2), enhance(2), . . .

Web: come from(13), begin in(9), haunt(8), start in(6), be(5), recur throughout(4), gild(3), begin from(3), be with(3), start at(3), . . .

0.24 “sand dune” MAKE_2

MTurk: be made of(17), consist of(6), contain(5), be composed of(4), be comprised of(4), be made from(3), be created from(3), be formed of(2), be formed by(2), be found in(2), . . .

Web: be(16), consist of(7), form in(6), move(5), be composed of(4), grow(4), have(3), accrete(3), be reported as(3), provide(3), . . .

0.23 “olive oil” FROM

MTurk: come from(19), be made from(11), be pressed from(6), be derived from(5), be made of(3), contain(3), originate from(3), taste like(2), run from(1), be found in(1), . . .

Web: be(74), include(31), be extracted from(22), be sold as(21), resemble(17), taste like(16), be obtained from(14), be made from(13), have(12), be pressed from(10), . . .

0.22 “history conference” ABOUT

MTurk: discuss(11), be about(8), focus on(5), involve(4), concern(4), deal with(3), talk about(3), be on(2), center on(2), concentrate on(2), . . .

Web: have(36), focus on(22), explore(15), make(12), look at(11), go down in(8), go in(8), consider(8), deal with(6), focus upon(6), . . .

0.22 “horse doctor” FOR

MTurk: treat(19), heal(9), care for(6), cure(5), tend to(4), medicate(4), work with(4), work on(3), fix(3), help(3), . . .

Web: ride(9), be(9), see(8), travel by(8), treat(7), have(7), use(6), dislike(5), come by(4), amputate(4), . . .

0.22 “salt lake” HAVE_1

MTurk: contain(19), have(16), include(4), be made of(3), be composed of(3), taste like(3), concentrate(2), taste of(2), be made from(2), hold(2), . . .

Web: be(71), fill(17), produce(15), have(12), contain(11), be located within(8), become(8), provide(6), afford(5), dissolve(5), . . .

0.22 “movement schisms” IN

MTurk: divide(5), involve(4), create(2), destroy(2), cause(2), shatter(2), come from(2), affect(2), split(1), occur with(1), . . .

Web: exist in(8), divide(7), beset(7), plague(6), wrack(4), rent(4), enervate(4), hobble(4), create(3), be mirrored in(3), . . .

0.21 “morning prayers” IN

MTurk: be said in(12), be recited in(4), occur in(4), start(3), be done in(2), be said(2), be said at(2), be recited at(2), relate to(2), occur at(2), . . .

Web: be held(16), be read(13), be recited(10), have(10), be offered(10), be said(9), be said in(6), go up(6), mark(6), include(6), . . .

0.21 "tire rim" HAVE$_2$

MTurk: surround(7), be on(5), hold(4), support(4), fit(2), be in(2), border(2), be of(2), be made for(2), cover(2), ...

Web: have(27), be used with(27), carry(24), support(17), hold(15), cause(13), fit(9), need(9), retain(8), require(8), ...

0.21 "soldier ant" BE

MTurk: act like(12), behave like(6), work like(6), be(5), act as(4), resemble(4), imitate(3), emulate(3), be called(2), fight like(2), ...

Web: bring(5), be(4), be killed by(3), act as(2), be turned into(2), be called(1), approach(1), go as(1).

0.21 "finger lakes" BE

MTurk: resemble(15), look like(12), be shaped like(7), emerge from(2), be thin like(2), flow in(2), remind of(2), appear like(2), have(1), contain(1), ...

Web: make up(15), look like(6), be(6), comprise(6), frame(4), stretch out like(3), precede(3), focus on(2), drain into(2), have(1), ...

0.21 "book requests" NOMINALIZATION:PRODUCT

MTurk: ask for(13), be made for(6), involve(5), be for(5), concern(4), refer to(3), pertain to(3), require(2), request(2), be about(2), ...

Web: be on(11), be for(4), publish(3), find(3), include(2), ask for(2), appear in(2), borrow(2), use(2), replicate(2), ...

0.21 "ceiling price" BE

MTurk: hit(5), be(5), reach(3), serve as(2), provide(2), form(2), go to(2), be at(2), stop at(2), approach(2), ...

Web: exceed(31), be(7), establish(7), set(7), be above(7), hit(6), act as(5), create(4), be below(4), go through(4), ...

0.21 "student problems" HAVE$_2$

MTurk: involve(6), concern(5), happen to(5), be experienced by(4), affect(4), belong to(4), be had by(4), be faced by(3), occur with(2), be caused by(2), ...

Web: require(507), affect(415), confront(209), prevent(163), allow(138), ask(116), interfere with(99), cause(95), engage(81), arise between(81), ...

0.21 "stone tools" MAKE$_2$

MTurk: be made of(15), be made from(7), work on(4), be used on(3), be created from(3), cut(2), shape(2), be manufactured from(2), be formed from(2), consist of(2), ...

Web: use in(52), cut(13), be made of(13), be used in(13), break(10), become(8), remove(8), be(7), move(5), bore into(5), ...

0.20 "hairpin turn" BE

MTurk: look like(16), resemble(14), be shaped like(5), mimic(4), be curved like(2), be abrupt like(1), turn as(1), ensnare like(1), be lik(1), be shaped as(1), ...

Web: promote(7), be(5), help for(3), resemble(2), look like(1), become(1).

0.20 "pressure cooker" USE

MTurk: use(16), cook with(12), utilize(10), create(5), involve(4), cook by(3), cook under(3), employ(3), require(3), generate(2), ...

Web: operate at(9), maintain(7), use(4), prevent(4), control(4), be(4), generate(3), increase in(3), provide(3), increase(2), ...

0.20 "color television" $HAVE_1$

MTurk: display(9), show(8), be in(7), display in(5), have(4), use(4), project in(3), broadcast in(3), transmit in(2), be found in(2), ...

Web: have(10), describe(4), be(4), reflect(3), invent(2), use(1), render(1), flash(1), create(1), match(1), ...

0.20 "gutter language" FROM

MTurk: come from(9), originate from(5), be used in(5), be from(5), resemble(4), belong in(3), be found in(2), stem from(2), emerge from(2), sound like(2), ...

Web: be of(3), be found in(2), savour of(2), settle in(2), embarrass(2), come from(1), belong in(1), rise from(1), escape from(1), be(1).

0.20 "glass eye" $MAKE_2$

MTurk: be made of(19), be composed of(11), be made from(6), be(4), be created from(3), consist of(3), be constructed of(2), resemble(2), look like(2), be created out of(2), ...

Web: need(35), be(24), wear(23), be of(18), look through(18), be hidden behind(17), require(16), be corrected with(16), have(12), be made of(11), ...

0.19 "budget speech" ABOUT

MTurk: be about(11), talk about(8), discuss(7), concern(5), involve(4), describe(4), refer to(4), address(3), explain(2), be on(1), ...

Web: affect(5), be on(3), be about(2), be taken in(2), be(2), be manifested through(1), be in(1), reflect(1), be delivered on(1).

0.19 "starvation diet" USE

MTurk: cause(9), use(6), lead to(5), involve(4), require(4), result in(4), come from(3), mimic(2), induce(2), go into(2), ...

Web: be(14), seem like(8), involve(7), prevent(6), tend toward(5), require(4), mean(4), mimic(3), advocate(3), be near(3), ...

0.19 "heat rash" $CAUSE_2$

MTurk: be caused by(26), come from(12), result from(5), be created by(4), stem from(3), originate from(2), be instigated by(1), be associated with(1), be exacerbated by(1), bespeak(1), ...

Web: start as(3), look like(3), be aggravated by(3), be associated with(2), be exacerbated by(2), occur in(2), be attended with(2), be known(2), be instigated by(1), come from(1), ...

0.18 "night flight" IN

MTurk: occur at(14), happen at(11), fly at(4), go at(3), be scheduled for(3), leave at(2), come at(2), be taken at(2), take off at(2), be scheduled during(2), ...

Web: leave at(28), arrive at(11), be(11), be at(11), arrive(7), operate at(5), leave(5), launch at(5), fly at(4), occur at(3), ...

0.18 "hermit crab" BE

MTurk: act like(10), behave like(7), resemble(7), be(6), live like(6), hide like(2), look like(2), emulate(2), stand for(1), isolate like(1), ...

Web: be(3), become(3), be like(2), carry around(2), clasp(2), investigate(1), include(1), follow(1), ones(1).

0.18 "water drop" $MAKE_2$

MTurk: be made of(16), contain(9), consist of(7), be(6), be composed of(6), be comprised of(3), be formed by(2), come from(2), have(1), emerge from(1), ...

Web: cause(23), be(13), fall into(13), make(10), bring back(8), suck in(8), contain(7), have(6), represent(6), grow in(6), ...

0.17 "artifact descriptions" NOMINALIZATION:PRODUCT

MTurk: describe(12), be about(5), define(5), explain(4), depict(3), be of(3), illuminate(3), tell about(3), catalog(2), discuss(2), ...

Web: be(5), deploy(5), be about(4), relate(2), allow(2), handle(1), be based on(1).

0.16 "student inventions" NOMINALIZATION:PATIENT

MTurk: be made by(19), be created by(13), come from(8), be invented by(5), be designed by(5), be developed by(3), involve(3), be produced by(3), originate from(1), commence from(1), ...

Web: allow(18), enable(6), be made by(4), be developed by(4), maximize(4), be performed by(4), be discussed by(3), reduce(3), move(2), show(2), ...

0.15 "milieu therapy" USE

MTurk: use(8), involve(4), utilize(4), modify(2), be based on(2), believe in(2), proceed in(1), occur in(1), be based upon(1), correct(1), ...

Web: alter(3), modify(2), improve(2), involve(1), be like(1).

0.15 "immigrant minority" $MAKE_2$

MTurk: be comprised of(9), include(7), be composed of(6), consist of(6), be made up of(6), contain(6), be made of(5), be(2), represent(2), involve(2), ...

Web: be(68), be viewed as(5), compete with(4), be drawn from(4), include(3), descend from(3), compete against(3), be expelled as(3), be composed of(2), consist of(2), ...

0.15 "phantom limb" BE

MTurk: be(10), feel like(10), be like(4), seem(3), act like(3), look like(2), resemble(2), appear(2), exist as(2), be sensed like(1), ...

Web: produce(4), be(1), be known as(1).

0.14 "machine translation" USE

MTurk: be done by(13), be performed by(6), come from(6), use(3), be produced by(3), be developed by(3), be completed by(3), be made by(3), require(3), originate from(2), ...

Web: use(9), make(8), have(6), allow(5), be(5), be called(4), take(4), run(4), be generated by(3), be produced by(3), ...

0.14 "alligator leather" FROM

MTurk: come from(14), be made from(13), resemble(3), be taken from(3), be made of(3), look like(2), feel like(2), be(2), be created from(2), be found on(2), ...

Web: look like(2), resemble(1), include(1).

0.14 "plant food" FOR

MTurk: feed(14), nourish(14), be for(8), sustain(5), be given to(5), be fed to(3), be used by(2), benefit(2), encourage(1), supply(1), ...

Web: come from(154), be derived from(74), kill(38), be(36), have(29), feed(26), operate(24), include(22), nourish(13), be used by(10), ...

0.13 "water wheel" USE

MTurk: use(8), move(6), be powered by(5), utilize(3), spin(2), work in(2), turn by(2), run on(2), rotate in(2), require(2), ...

Web: receive(44), raise(39), lift(32), pump(28), be turned by(26), draw(24), spill(20), be powered by(13), bring(13), drive(11), ...

0.13 "animal attack" NOMINALIZATION:ACT

MTurk: be made by(10), be caused by(9), come from(7), involve(6), be initiated by(4), be done by(4), be performed by(3), be blamed on(2), be started by(2), be generated by(2), ...

Web: make(5), appear in(4), leave(4), come from(2), provoke(2), deprive(2), be prevented by(2), cause(2), involve(1), be performed by(1), ...

0.12 "job tension" CAUSE$_2$

MTurk: be caused by(14), come from(8), result from(6), be related to(5), be created by(5), be due to(3), involve(3), be associated with(2), be found in(2), emerge from(2), ...

Web: make(14), come from(5), cause(4), exist in(4), come with(2), harm(2), stir(2), go with(2), be(2), do(2), ...

0.12 "radio communication" USE

MTurk: use(7), come from(7), need(3), come by(3), be broadcast by(2), be made by(2), be carried by(2), emanate from(2), be transmitted by(2), pass through(2), ...

Web: provide(29), include(28), be(27), be broadcast by(14), bring(13), be transmitted over(10), own(8), run over(7), involve(6), use(6), ...

0.12 "lace handkerchief" HAVE$_1$

MTurk: be made of(18), be made from(7), be composed of(4), contain(4), consist of(4), be trimmed with(3), be decorated with(3), resemble(3), use(3), be(3), ...

Web: feature(1), be trimmed with(1).

0.12 "evening hours" IN

MTurk: occur in(13), be in(9), occur during(5), include(2), involve(2), happen during(2), be of(2), refer to(2), pertain to(2), happen in(2), ...

Web: include(554), be(29), extend into(29), be in(11), be during(10), stretch into(9), involve(6), run into(6), require(6), start(5), ...

0.12 “wood shavings” FROM

MTurk: come from(17), be made of(11), be made from(6), be composed of(4), contain(3), be produced from(3), be comprised of(3), be derived from(2), be created by(2), originate from(1), . . .

Web: contain(5), be obtained from(2), make(1).

0.12 “abortion problem” ABOUT

MTurk: involve(9), concern(6), be caused by(4), come from(3), surround(3), occur in(3), stem from(3), be related to(3), relate to(3), occur with(2), . . .

Web: cause(17), include(13), follow(9), be associated with(7), arise from(5), require(5), be caused by(4), result from(4), occur during(4), be created by(3), . . .

0.11 “parent organization” BE

MTurk: be made up of(6), include(5), be composed of(4), involve(4), consist of(4), contain(4), be for(3), be run by(2), act like(2), concern(2), . . .

Web: help(366), work with(238), provide(207), assist(200), support(160), serve(154), be(136), represent(93), bring(87), educate(81), . . .

0.10 “bull ring” FOR

MTurk: contain(12), hold(4), be worn by(3), have(3), house(3), be for(3), resemble(2), exhibit(2), feature(2), surround(2), . . .

Web: resemble(7), be(7), be in(5), look like(5), grip(5), face(4), form(4), comprise(3), produce(2), belong on(2), . . .

0.10 “people power” HAVE_2

MTurk: come from(12), belong to(5), be used by(3), be generated by(3), be derived from(2), involve(2), be produced by(2), be given to(2), be possessed by(2), reside with(1), . . .

Web: enable(108), allow(87), be(68), lead(66), make(64), bring(54), force(54), affect(48), attract(48), oppress(47), . . .

0.10 “heart design” BE

MTurk: look like(10), resemble(8), be shaped like(8), contain(6), include(5), incorporate(3), use(3), involve(3), feature(2), be made from(2), . . .

Web: be(60), be at(59), capture(45), win(34), warm(28), include(21), lie at(19), form(18), touch(12), drive with(12), . . .

0.10 “concrete desert” MAKE_2

MTurk: be made of(14), be made from(5), be composed of(5), resemble(5), consist of(5), contain(4), look like(3), be(2), be in(2), be covered in(1), . . .

Web: be covered in(2), be(1).

0.09 “abortion vote” ABOUT

MTurk: be about(7), involve(6), concern(6), allow(3), pertain to(3), support(3), affect(3), decide(2), decide on(2), approve(2), . . .

Web: legalize(6), further(4), outlaw(4), allow(2), criminalize(2), permit(1).

0.09 “office friendships” IN

MTurk: start in(6), begin at(5), come from(4), form in(4), be made in(3), occur in(3), begin in(3), be in(3), be formed in(3), occur at(3), . . .

Web: improve by(4), extend outside(3), elevate(2), be cemented by(2), extend beyond(2), make(2), share(2), start in(1), be made in(1), transcend(1), . . .

0.09 “tax law” ABOUT

MTurk: be about(9), involve(7), govern(4), describe(4), regulate(3), be related to(3), pertain to(3), concern(3), deal with(2), relate to(2), . . .

Web: impose(476), reduce(94), raise(88), pay(62), levy(61), increase(41), place(31), require(30), establish(30), allow(30), . . .

0.09 “moth hole” CAUSE$_2$

MTurk: be made by(16), be created by(8), be caused by(6), be eaten by(5), be produced by(4), come from(3), be generated by(2), result from(2), be perpetrated by(1), protect(1), . . .

Web: let(3), be caused by(1), be(1).

0.09 “faith cure” USE

MTurk: rely on(6), come from(5), be caused by(5), require(4), depend on(3), rely upon(2), depend upon(2), occur because of(2), be due to(2), use(2), . . .

Web: be wrought by(7), rest in(4), be activated by(3), be obtained by(3), establish(3), ignore(3), be performed by(3), require(2), come from(1), be produced through(1), . . .

0.08 “police intervention” NOMINALIZATION:ACT

MTurk: involve(12), be done by(6), come from(5), require(5), be performed by(5), be made by(3), include(2), be conducted by(2), occur with(1), occur through(1), . . .

Web: bring(16), send(6), allow(6), involve(2), be implemented by(2), do iniohe(2), include(1), divert(1), assist(1), force(1).

0.08 “energy emergency” ABOUT

MTurk: involve(12), concern(6), relate to(4), be due to(3), be caused by(3), be about(3), require(2), emerge from(2), refer to(2), come from(2), . . .

Web: demand(16), call(7), require(3), give(3), affect(2), deplete(2), tax(2), call for(2), bed(1), allow(1).

0.08 “hydrogen bomb” USE

MTurk: contain(9), use(8), be made from(5), be composed of(5), be made of(5), fuse(3), utilize(3), consist of(3), have(2), involve(2), . . .

Web: squeeze(5), fuse(4), detonate(4), develop(3), combine(3), run on(2), be known as(2), make(2), be(2), oppose(2), . . .

0.08 “summer months” IN

MTurk: occur during(13), be in(9), occur in(8), happen in(7), make up(3), be found in(3), comprise(3), include(2), be during(2), represent(2), . . .

Web: be(47), follow(24), include(17), end(8), make up(7), constitute(7), start(7), masquerade as(6), be considered(5), be employed during(5), . . .

0.07 "family antiques" HAVE$_2$

MTurk: belong to(17), be owned by(11), be passed down through(5), come from(4), be kept in(2), pass through(2), stay in(2), be kept by(2), be associated with(2), be in(1), ...

Web: be in(77), be passed down through(6), be kept in(4), be handed down through(4), belong in(4), come from(3), be honored in(2), be passed down in(1), be compiled since(1), be passed down(1), ...

0.07 "kennel puppies" FROM

MTurk: live in(17), come from(9), be born in(8), be raised in(6), be bred in(6), be found in(3), be from(3), be adopted from(2), be housed in(2), inhabit(2), ...

Web: pee in(110), be kept in(12), be raised in(11), be in(6), come from(4), love(4), live in(3), leave(3), be offered by(3), die in(2), ...

0.07 "murder charge" BE

MTurk: involve(14), accuse of(4), come from(3), concern(3), accuse(2), emerge from(2), relate to(2), consist of(2), imply(2), include(1), ...

Web: include(228), attempt(34), be(18), commit(8), range from(8), instigate(8), be attempted(4), solicit for(3), be reduced from(3), involve(2), ...

0.06 "summer travels" IN

MTurk: occur during(12), happen in(10), occur in(8), begin in(3), involve(3), be in(3), be taken in(3), occur(2), be done in(2), happen during(2), ...

Web: begin in(3), take(3), increase in(3), be executed during(2), work for(2), be presented in(2), begin(2), be encountered during(2), run from(1), be embarked on(1), ...

0.06 "pet spray" FOR

MTurk: be used on(9), be for(6), be used for(6), be applied to(5), be made for(3), protect(2), medicate(2), be sprayed on(2), repel(2), be applied on(2), ...

Web: interrupt(43), keep(18), deter(12), be used on(4), cleanse(3), condition(3), hydrate(2), harm(2), make(2), train(2), ...

0.06 "infant colonies" BE

MTurk: contain(7), be made of(4), be composed of(3), resemble(3), be like(3), be(2), be comprised of(2), be made up of(2), support(2), consist of(2), ...

Web: have(1), affect(1).

0.06 "water mark" CAUSE$_2$

MTurk: be made by(5), be created by(5), indicate(4), be made from(4), come from(4), be caused by(4), look like(3), be left by(3), be made with(3), show(1), ...

Web: be covered by(16), be(9), be removed with(8), be removed by(6), be inundated by(5), modify(5), show(4), have(4), be on(3), be washed out with(3), ...

0.06 "dream analysis" NOMINALIZATION:ACT

MTurk: explain(8), interpret(7), involve(5), study(4), discuss(3), translate(3), be about(3), investigate(2), examine(2), refer to(2), ...

Web: refer(7), have(6), be compared with(5), include(4), be given of(4), address(4), present(3), contain(2), embody(2), bring(2), ...

0.06 "student power" $HAVE_2$

MTurk: come from(10), belong to(5), be held by(4), be used by(4), be possessed by(4), involve(3), be created by(3), be wielded by(2), be demonstrated by(2), be exercised by(2), . . .

Web: enable(34), be(11), draw(10), keep(10), admit(8), exist between(8), take(8), arrest(7), make(7), allow(7), . . .

0.06 "subject deletion" NOMINALIZATION:ACT

MTurk: remove(12), eliminate(8), involve(5), include(3), erase(3), delete(2), affect(2), destroy(2), wipe out(2), be of(1), . . .

Web: be(21), delete(10), be found in(2), be on(2), drop(2), affect(1), limit(1), permit(1), be about(1).

0.06 "nose drops" FOR

MTurk: be used in(4), go in(4), be placed in(4), clear(4), medicate(3), go into(3), help(3), heal(3), come from(3), be applied in(2), . . .

Web: run down(6), be inhaled through(3), be suspended from(3), pay thru(3), hang from(2), fell from(2), go down(2), get into(2), be used in(1), hang at(1), . . .

0.05 "adolescent turmoil" IN

MTurk: be experienced by(7), affect(7), occur in(6), involve(6), happen to(5), be caused by(3), surround(2), afflict(2), relate to(2), befall(2), . . .

Web: accompany(2), surround(1).

0.05 "cash basis" BE

MTurk: involve(7), use(6), be based on(5), require(4), include(3), come from(3), depend on(3), be made of(2), consist of(2), accept(2), . . .

Web: recognise(62), be(15), include(13), recognize(12), generate(11), measure(9), mean that(7), affect(7), charge(4), be in(4), . . .

0.05 "student critiques" NOMINALIZATION:PRODUCT

MTurk: be made by(6), be written by(4), be given by(4), come from(3), involve(2), be done by(2), be made about(2), be developed by(2), be created by(2), concern(2), . . .

Web: allow(14), help(9), engage(4), be sent between(4), require(4), encourage(3), present(3), provide(3), motivate(3), be written by(2), . . .

0.05 "birth control" NOMINALIZATION:ACT

MTurk: prevent(11), regulate(7), stop(7), limit(5), control(2), inhibit(2), discourage(2), involve(2), reduce(2), concern(2), . . .

Web: give(92), acne(10), have(8), be held after(7), prevent(5), occur after(5), result in(4), be(4), rely on(4), make for(3), . . .

0.05 "oil crisis" ABOUT

MTurk: involve(9), be caused by(8), be related to(3), be about(2), revolve around(2), be precipitated by(2), relate to(2), depend on(2), concern(2), need(1), . . .

Web: bear upon(12), be known as(11), follow(10), arise from(7), originate in(7), result from(6), impact(5), send(5), push(4), be(4), . . .

0.05 "weekend boredom" IN

MTurk: occur on(15), happen during(9), happen on(9), occur during(5), come on(3), plague(1), happen along(1), coincide with(1), happen through(1), occur through(1), ...

Web: plague(2).

0.05 "plum wine" FROM

MTurk: be made from(19), contain(7), taste like(7), come from(7), taste of(2), include(2), be manufactured from(2), be made of(2), need(2), be produced from(2), ...

Web: have(14), taste of(9), offer(9), show(6), take(6), evoke(5), be(4), be packed with(4), resemble(3), be loaded with(3), ...

0.05 "steel helmet" $MAKE_2$

MTurk: be made of(17), be made from(7), contain(6), be composed of(5), be manufactured from(4), be made out of(3), be comprised of(3), be fashioned from(3), be constructed from(2), consist of(2), ...

Web: be of(19), sparkle like(2), go under(2), be like(1), replace(1), fit inside(1).

0.04 "warrior castle" $MAKE_2$

MTurk: house(11), be lived in by(5), be inhabited by(5), belong to(4), contain(4), be owned by(3), protect(2), hold(2), be occupied by(2), support(2), ...

Web: be defended by(3), be thronged with(3), produce(1), allow(1).

0.04 "smoke signals" USE

MTurk: be made of(11), use(8), be made from(5), consist of(5), originate from(2), be composed of(2), resemble(2), emerge from(2), be comprised of(2), be made up of(2), ...

Web: provide(5), raise(5), contain(4), be(4), be corrected by(4), be altered by(3), distinguish between(2), emerge as(2), be obscured by(2), include(1), ...

0.04 "tape measure" BE

MTurk: be made of(8), use(5), look like(5), be(3), be on(3), resemble(3), be made from(2), be in(2), consist of(2), have(1), ...

Web: reduce(22), cut(18), have(9), prevent(8), make(7), eliminate(5), protect(4), include(4), retract(4), be(3), ...

0.04 "onion tears" $CAUSE_2$

MTurk: be caused by(18), come from(7), result from(5), be created by(3), react to(2), derive from(2), be induced by(2), be generated by(2), come with(1), be increased by(1), ...

Web: come with(6), chop(4), live in(3).

0.04 "surface tension" IN

MTurk: be on(5), occur on(5), occur at(4), happen at(3), exist on(2), affect(2), involve(2), reside on(2), happen on(2), be found on(2), ...

Web: lie beneath(69), simmer beneath(44), lurk beneath(30), lay beneath(23), simmer under(22), cause(22), be under(20), be below(18), simmer below(17), clean(16), ...

0.03 "voice vote" USE

MTurk: be made by(9), use(6), be given by(3), be cast using(3), come from(3), originate from(2), count(2), come by(2), be conducted by(2), be counted by(2), ...

Web: be(13), have(9), give(9), represent(9), be rejected by(5), be by(4), muffle(3), carry by(3), provide(3), pass by(3), ...

0.03 "air pressure" CAUSE$_2$

MTurk: come from(9), be caused by(7), be due to(4), be made by(4), use(3), involve(3), be measured in(3), result from(2), be related to(2), pertain to(2), ...

Web: draw(109), cause(97), force(67), pull(41), prevent(34), be(27), drive(23), suck(19), contaminate(18), push(15), ...

0.03 "sugar cube" MAKE$_2$

MTurk: be made of(20), contain(9), consist of(9), be made from(7), be composed of(5), be(4), be comprised of(3), be formed from(3), taste like(2), look like(1), ...

Web: be dusted with(8), compose of(8), drip on(2), be dusted in(2), look like(1), contain(1), be steeped in(1), be covered in(1).

0.03 "basketball season" FOR

MTurk: include(8), involve(5), feature(5), contain(4), have(3), be for(3), center around(2), promote(2), focus on(2), play(1), ...

Web: be(6), play(3), be followed by(3), lift(2), care about(1), mean(1), come after(1).

0.03 "party members" HAVE$_2$

MTurk: belong to(13), attend(5), join(4), go to(4), represent(2), form(2), comprise(2), be in(2), work for(2), participate in(2), ...

Web: be(3555), leave(190), represent(186), join(165), attend(74), contract(72), become(62), switch(48), quit(35), support(34), ...

0.02 "pot high" CAUSE$_2$

MTurk: come from(12), be caused by(8), result from(6), be derived from(4), emerge from(3), require(2), be created by(2), be induced by(2), be due to(2), involve(2), ...

Web: extend over(3), take(2), score(2), require(1), lift(1), be kept in(1), make(1), melt(1), be planted in(1), keep(1).

0.02 "land reclamation" NOMINALIZATION:ACT

MTurk: involve(6), take back(4), recover(4), restore(4), recapture(2), renew(2), create(2), reclaim(2), affect(2), include(1), ...

Web: require(13), withdraw(8), prevent(5), own(5), provide(4), merge in(4), return(4), make(3), manage(3), occur on(2), ...

0.02 "pole height" HAVE$_2$

MTurk: measure(4), describe(3), define(3), be reached by(2), reach(2), refer to(2), come from(2), be set for(1), originate from(1), belong to(1), ...

Web: allow(6), cause(3), vary from(2), be set for(1).

0.02 "pet theory" BE

MTurk: involve(6), concern(6), be about(6), pertain to(3), apply to(2), be(2), refer to(2), relate to(2), be like(2), revolve around(1), ...

Web: satisfy(4), line up with(3), revolve around(1), challenge(1).

0.01 "government employment" IN

MTurk: involve(4), be provided by(4), be sponsored by(3), be with(3), benefit(3), be made by(2), be within(2), come from(2), be offered by(2), assist(1), ...

Web: include(12), be adopted by(12), be generated by(6), be registered with(4), be promoted by(4), ask(4), be(4), be slowed down by(4), reduce(4), be endorsed by(4), ...

0.00 "apple core" HAVE_2

MTurk: come from(13), be inside(5), be in(4), be found in(3), be within(3), be located in(2), be from(2), fill(1), belong to(1), be derived from(1), ...

Web: win(2), be used in(2), be removed from(2), show up in(1), be(1), run(1).

0.00 "arms budget" FOR

MTurk: pay for(7), be for(6), buy(5), concern(4), be meant for(3), include(3), purchase(2), be used for(2), deal with(2), cover(2), ...

Web: be(4), play(3), cost(2), back(1), give(1), permit(1), go for(1).

0.00 "beard trim" NOMINALIZATION:PRODUCT

MTurk: shorten(8), cut(6), involve(4), happen to(3), neaten(3), affect(3), beautify(2), trim(2), be done to(2), be for(2), ...

Web: .

0.00 "bird reproduction" NOMINALIZATION:ACT

MTurk: involve(7), be done by(4), resemble(3), be caused by(3), reproduce(3), occur in(2), represent(2), make(2), be of(2), be(2), ...

Web: be imposed on(1), be adorned with(1), show(1).

0.00 "blanket excuse" BE

MTurk: cover like(9), resemble(5), be like(5), act like(4), act as(2), work like(2), be used as(1), be analogous to(1), behave like(1), seem like(1), ...

Web: .

0.00 "canine companion" BE

MTurk: be(21), resemble(3), belong to(2), bark like(2), look like(2), be classified as(2), be from(2), be born(2), be made from(1), be of(1), ...

Web: cheer on(3), play with(2).

0.00 "cigarette war" ABOUT

MTurk: involve(12), concern(5), be about(5), be over(3), revolve around(2), be fought over(2), ban(2), focus on(2), be caused by(2), be caused for(1), ...

Web: democratise(1), give(1), get(1).

0.00 “coffee nerves” CAUSE$_2$

MTurk: be caused by(13), come from(8), result from(5), need(3), be affected by(2), be blamed on(2), emerge from(2), be from(2), be irritated by(1), occur because of(1), ...

Web: .

0.00 “communist tenet” IN

MTurk: be believed by(10), be held by(6), be written by(5), support(4), belong to(3), be followed by(3), be made by(3), come from(3), emerge from(2), define(2), ...

Web: .

0.00 “coriander curry” HAVE$_1$

MTurk: contain(18), include(10), be made from(8), have(7), be made with(7), come from(4), taste like(3), be flavored with(3), use(3), be garnished with(2), ...

Web: roast(1).

0.00 “daisy chains” MAKE$_2$

MTurk: be made of(12), be made from(9), contain(5), consist of(5), be comprised of(4), look like(4), use(3), link(3), be created with(2), resemble(2), ...

Web: be(4), be stabilised by(4).

0.00 “disaster flick” ABOUT

MTurk: be about(12), involve(6), portray(6), describe(5), depict(3), show(3), concern(3), contain(2), be written about(1), revolve around(1), ...

Web: be saved from(2), merge(2).

0.00 “enemy strength” HAVE$_2$

MTurk: belong to(6), be possessed by(6), be owned by(3), be held by(2), be of(2), describe(2), be attributed to(2), be posessed by(2), evaluate(2), empower(2), ...

Web: destroy(9), bring(9), be turned against(6), deter(6), crush(6), be(4), bear down(4), overthrow(3), break(3), be used against(3), ...

0.00 “factory rejects” NOMINALIZATION:PATIENT

MTurk: come from(14), involve(4), be made by(4), be made in(4), be rejected in(2), be produced by(2), be discarded by(2), be from(2), belong to(1), occur in(1), ...

Web: .

0.00 “financing dilemma” ABOUT

MTurk: involve(9), be caused by(6), concern(6), be related to(4), come from(4), affect(4), relate to(3), result from(3), be about(3), be due to(2), ...

Web: .

0.00 “handlebar mustache” BE

MTurk: look like(20), resemble(17), be shaped like(6), appear(2), appear like(2), be used as(1), clone(1), be formed like(1), suggest(1), mimic(1), ...

Web: twist into(1), hang in(1).

0.00 "heart massage" NOMINALIZATION:ACT

MTurk: stimulate(9), help(5), revive(4), restart(4), be performed on(2), be applied to(2), involve(2), palpate(2), treat(2), start(2), . . .

Web: connect(8), nourish(5), bruise(3), reach(2), get into(2), slow(2), win(1), delight(1), work(1), be affirmed in(1).

0.00 "jungle exploration" NOMINALIZATION:ACT

MTurk: occur in(8), happen in(7), involve(6), go into(6), be in(5), examine(3), enter(3), travel through(2), be done in(2), explore(2), . . .

Web: figure(1), ascend from(1).

0.00 "language riots" ABOUT

MTurk: be caused by(8), involve(6), be about(5), include(3), concern(3), be inspired by(2), happen because of(2), come from(2), result from(2), occur with(1), . . .

Web: follow(2).

0.00 "laugh wrinkles" $\mathbf{CAUSE_2}$

MTurk: be caused by(17), come from(15), be created by(4), result from(4), bespeak(1), be derived from(1), show(1), emanate from(1), be casued by(1), stem from(1), . . .

Web: add(1), make(1).

0.00 "lemon peel" $\mathbf{HAVE_2}$

MTurk: come from(19), surround(6), protect(5), cover(5), be found on(4), smell like(4), come off(4), be taken from(3), be made from(2), be from(2), . . .

Web: .

0.00 "liquor orders" NOMINALIZATION:PRODUCT

MTurk: request(11), contain(8), be for(7), include(4), purchase(4), ask for(4), require(3), demand(2), involve(2), list(2), . . .

Web: indulge in(3), intoxicate(2), permit(2), require intoxicating(1).

0.00 "midnight snack" IN

MTurk: be eaten at(19), be consumed at(11), happen at(7), occur at(5), occur around(2), be eaten around(2), be made at(2), be snuck at(2), be taken at(2), be consumed around(1), . . .

Web: .

0.00 "morphology lecture" ABOUT

MTurk: be about(13), explain(6), discuss(5), describe(5), concern(4), involve(3), teach(3), cover(3), include(2), deal with(2), . . .

Web: .

0.00 "pedal extremeties" BE

MTurk: be(6), include(3), use(3), push(3), resemble(2), move(2), be used to(2), look like(2), contain(2), be situated on(2), . . .

Web: .

0.00 “pork suet” FROM

MTurk: come from(11), be made from(10), be made of(8), contain(5), be derived from(3), taste like(3), be rendered from(3), consist of(3), originate from(2), be composed of(2), ...

Web: .

0.00 “rice paper” FROM

MTurk: be made of(15), be composed of(8), be made from(6), contain(5), be created from(4), consist of(4), be produced from(3), come from(3), be made out of(2), be comprised of(2), ...

Web: be(7), domesticate(4), range from(2), affirm(2), have(2), be eaten with(1), be weighted down with(1), be called(1), precoated(1).

0.00 “ship landing” NOMINALIZATION:ACT

MTurk: be used by(6), be for(4), involve(3), be made by(3), service(3), be meant for(2), hold(2), be done by(2), dock(2), receive(2), ...

Web: be made in(3), bring(2), cause(2), look like(2), break(2), maneuver(2), allow(2), be with(1), damage(1), keep(1).

0.00 “staff attempts” NOMINALIZATION:ACT

MTurk: be made by(18), come from(5), be done by(4), be performed by(3), originate from(2), be initiated by(2), be undertaken by(2), be from(2), be tried by(2), be engendered by(1), ...

Web: upset(2), be made for(1).

0.00 “testtube baby” FROM

MTurk: be conceived in(13), come from(10), be created in(8), be made in(7), begin in(4), be started in(4), originate in(4), be fertilized in(4), originate from(2), start in(2), ...

Web: .

0.00 “tobacco ash” FROM

MTurk: come from(21), derive from(5), be produced from(4), contain(4), be made of(3), be from(3), be made from(2), be derived from(2), be(2), emerge from(2), ...

Web: fall from(5), be mixed with(2).

0.00 “vapor lock” CAUSE_2

MTurk: be caused by(11), result from(5), contain(4), involve(2), keep out(2), control(2), come from(2), occur because of(1), keep inside(1), inhibit(1), ...

Web: .

0.00 “wastebasket category” BE

MTurk: belong in(6), include(4), go in(3), describe(2), come from(2), be about(2), fill(1), belong to(1), be named(1), be put into(1), ...

Web: .

E.2 Comparison by Noun Compound: Using The First Verb Only

0.99 "flu virus" CAUSE$_1$

MTurk: cause(17), be(1), carry(1), involve(1), come from(1).

Web: cause(906), produce(21), give(20), differentiate(17), be(16), have(13), include(11), spread(7), mimic(7), trigger(6), ...

0.99 "disease germ" CAUSE$_1$

MTurk: cause(18), carry(3), spread(2), produce(1), generate(1).

Web: cause(919), produce(63), spread(37), carry(20), propagate(9), transmit(7), be(7), create(7), bring(5), give(4), ...

0.97 "student friends" BE

MTurk: be(15), be had by(2), meet as(1), help(1), be considered(1), come from(1), be made by(1), exist as(1), comrise(1), matriculate as(1).

Web: be(1250), have(34), support(33), know(31), teach(22), give(21), meet as(19), help(16), guide(15), pose as(12), ...

0.97 "fruit tree" HAVE$_1$

MTurk: bear(18), produce(7), have(2), bare(1), grow(1), bloom(1).

Web: bear(2113), produce(1215), have(337), bore(269), yield(201), bring(157), provide(155), bringeth(140), give(135), grow(74), ...

0.96 "women professors" BE

MTurk: be(21), teach(1), look like(1), resemble(1).

Web: be(251), teach(46), study(38), specialize in(26), appear as(13), think(10), take(10), work with(9), think that(9), research(9), ...

0.95 "citizen soldier" BE

MTurk: be(19), protect(1), be classified as(1), represent(1), be selected from(1), live as(1), come from(1).

Web: be(185), become(44), be treated as(14), save(14), dehumanize(11), shoot(10), rob(6), murder(5), fire upon(5), be drafted from(5), ...

0.95 "cooking utensils" FOR

MTurk: be used for(15), be used in(7), facilitate(1), be found in(1), be made for(1), aid in(1), aid(1), underscore(1).

Web: be used for(43), be used in(11), make(6), be suited for(5), replace(3), facilitate(2), turn(2), be used during(2), keep(2), be for(1), ...

0.94 "music box" MAKE$_1$

MTurk: play(15), produce(3), make(3), include(1), contain(1), be made for(1), deliver(1).

Web: play(104), make(34), produce(18), have(16), provide(14), be(13), contain(9), access(8), say(7), store(6), ...

0.94 "desert rat" IN

MTurk: live in(21), be located in(1), be found in(1), inhabitates(1), dwell in(1), come from(1).

Web: live in(16), occur in(4), do(3), inhabit(2), inhibit(1), love(1), survive in(1).

0.93 "city folk" IN

MTurk: live in(17), come from(3), reside in(3), emerge from(1).

Web: live in(245), run(26), be in(22), leave(22), move from(22), come from(19), work in(17), flee(17), make(16), populate(13), ...

0.93 "accident weather" CAUSE_1

MTurk: cause(13), promote(2), be condusive to(1), lead to(1), occur by(1), result in(1), precipitate(1), be conducive to(1), foster(1), happen by(1), ...

Web: cause(25), provoke(3), be(3), include(2), end in(2), exist during(2), be proven with(1).

0.93 "blood donor" NOMINALIZATION:AGENT

MTurk: give(25), donate(5).

Web: give(653), donate(395), receive(74), sell(41), provide(39), supply(17), be(13), match(11), contribute(10), mean(9), ...

0.93 "city wall" HAVE_2

MTurk: surround(20), protect(2), enclose(2), contain(1), encircle(1), be in(1), border around(1), surronds(1).

Web: surround(708), encircle(203), protect(191), divide(176), enclose(72), separate(49), ring(41), be(34), encompass(25), defend(25), ...

0.92 "gas stove" USE

MTurk: use(15), run on(4), burn(2), emit(2), run by(1), be powered by(1), heat by(1), exumes(1).

Web: use(98), run on(36), burn(33), be(25), be heated by(10), work with(7), be used with(7), leak(6), need(6), be run with(6), ...

0.91 "mail sorter" NOMINALIZATION:AGENT

MTurk: sort(12), process(3), handle(2), organize(2), arrange(1), divide(1), work with(1), tend to(1), deal with(1), look through(1).

Web: sort(25), sweep(4), separate(1), deliver(1), result in(1).

0.90 "target structure" BE

MTurk: be(10), have(3), include(2), hold(1), represent(1), contain(1), become(1), aim at(1), resemble(1), uphold(1), ...

Web: be(513), serve as(68), make(45), become(34), represent(21), form(18), provide(18), constitute(13), host(13), mimic(12), ...

0.90 "honey bee" MAKE_1

MTurk: make(16), produce(7), create(2), provide(1), manufacture(1), store(1).

Web: make(292), produce(189), gather(104), have(69), collect(57), suck(39), extract(19), drink(16), bring(15), carry(13), ...

0.89 “song bird” MAKE$_1$

MTurk: sing(19), make(2), produce(1), warble(1), love(1), erupt in(1).

Web: sing(264), learn(80), have(69), hear(41), be(27), lose(18), continue in(18), know(15), develop(11), give(11), ...

0.88 “snow blindness” CAUSE$_2$

MTurk: be caused by(22), come from(3), be created by(1), exemplify(1).

Web: be caused by(2), be like(1).

0.88 “servant girl” BE

MTurk: be(10), work as(8), act as(5), be used as(1), be employed as(1), become(1), exemplify(1), come from(1).

Web: be(193), work as(119), become(78), live as(20), require(16), live(10), be employed as(9), grow up as(9), go as(9), toil as(8), ...

0.88 “novelty item” BE

MTurk: be(11), display(1), be considered(1), be sold as(1), represent(1), become(1), be loved for(1), have(1), resemble(1), be regarded as(1), ...

Web: be(14), include(4), display(3), range from(3), possess(2), appear as(2), be considered(1), be imroduced as(1), seem like(1).

0.86 “field mouse” IN

MTurk: live in(19), inhabit(2), be in(1), roam around(1), originate in(1), dwell in(1), come from(1).

Web: live in(38), ravage(10), be(8), suggest that(8), have(6), burrow in(5), infest(5), occur in(4), extend(4), be deprived of(3), ...

0.86 “hand brake” USE

MTurk: be operated by(9), be applied by(3), operate by(3), require(3), be used by(2), use(1), be applyed by(1), be operated with(1), go in(1), need(1), ...

Web: be operated by(112), be applied by(14), be(6), be controlled by(6), be set by(5), be released by(5), require that(4), require(3), be of(3), secure(3), ...

0.86 “concussion force” CAUSE$_1$

MTurk: cause(17), create(3), result in(2), produce(1), be found in(1), lead to(1), induce(1), form(1), be produced by(1).

Web: cause(25), produce(15), suffer(2), make(2), be termed(1).

0.83 “sex scandal” ABOUT

MTurk: involve(12), be about(4), be caused by(2), have(1), center around(1), relate to(1), deal with(1), be made because of(1), concern(1), be concerned with(1), ...

Web: involve(46), go beyond(21), college(3), revolve around(3), turn into(3), be(3), have(2), originate as(2), include(2), make(2), ...

0.83 "collie dog" BE

MTurk: be(10), be called(2), look like(1), come from(1), resemble(1), belong to(1), emerge from(1), be named(1), be implied to(1), be bred from(1), . . .

Web: be(24), look like(14), resemble(8), be border(5), feature(3), come from(2), tend(2), be bearded(1), include(1), betoken(1), . . .

0.82 "cigarette burn" $CAUSE_2$

MTurk: be caused by(13), come from(6), be made by(2), resemble(1), be done with(1), be left by(1), result from(1).

Web: be caused by(6), come from(5), be from(4), consume(2), be made by(1), resemble(1), be inflicted by(1), suggest(1).

0.80 "film cutter" NOMINALIZATION:AGENT

MTurk: cut(14), slice(5), work with(3), edit(2), trim(1), be used for(1), be used on(1), run through(1), splice(1), come from(1), . . .

Web: cut(20), see(7), divide(5), chop up(3), conform(3), separate(2), perforate(2), sever(2), act in(1), be reciprocated across(1), . . .

0.80 "peer judgments" NOMINALIZATION:PRODUCT

MTurk: be made by(13), come from(8), emanate(1), be handed down from(1), refer to(1), be held by(1), be given by(1), be passed by(1).

Web: be made by(3), be had against(1).

0.80 "farm boy" FROM

MTurk: live on(14), work on(6), come from(4), work in(3), reside on(1).

Web: live on(175), grow up on(103), work on(72), leave(67), remain on(13), be raised on(12), go from(11), be reared on(9), come from(8), live in(8), . . .

0.78 "deficiency disease" $CAUSE_2$

MTurk: be caused by(12), come from(2), result in(1), stem from(1), be associated with(1), have(1), involve(1), result from(1), cause(1), arise from(1), . . .

Web: be caused by(253), result from(203), be characterized by(81), have(69), cause(59), be associated with(56), involve(45), result in(40), be(39), arise from(34), . . .

0.77 "pet families" $HAVE_1$

MTurk: have(9), own(5), be made of(2), like(1), care for(1), keep(1), enjoy(1), consist of(1), adopt(1), love(1), . . .

Web: have(187), adopt(84), lose(77), love(46), want(43), own(38), bring(35), keep(28), include(24), take(24), . . .

0.77 "head noun" BE

MTurk: be(6), involve(2), refer to(2), be at(2), be found in(2), occur at(1), form(1), act as(1), describe(1), represent(1), . . .

Web: be(149), form(24), function as(16), occur as(14), appear as(13), modify(10), act as(7), serve as(7), precede(5), be analysed as(4), . . .

0.73 "sap tree" MAKE_1

MTurk: produce(12), have(4), supply(2), give(2), exude(1), make(1), create(1), ooze(1), grow(1), ooze out(1).

Web: produce(78), force(52), have(42), drip(33), exude(32), ooze(20), cause(18), get(17), secrete(16), yield(15), ...

0.72 "grain alcohol" FROM

MTurk: be made from(10), come from(4), be produced from(3), be distilled from(2), be made of(2), contain(2), be derived from(1), be fermented from(1), incorporate(1), be extracted from(1).

Web: be made from(13), be produced from(11), be manufactured from(6), be derived from(5), require(4), crystallize in(3), raise(3), be called(1), mean(1), produce from(1), ...

0.70 "beehive hairdo" BE

MTurk: look like(11), resemble(9), be shaped like(3), emulate(1), rise like(1).

Web: look like(4), be described as(4), resemble(2), be(1).

0.69 "tear gas" CAUSE_1

MTurk: cause(15), create(2), produce(1), generate(1), shrades(1).

Web: cause(13), choke as(8), be like(6), include(4), tamponades(4), be(4), be made from(3), induce(3), trigger(2), bring(2), ...

0.69 "growth hormone" CAUSE_1

MTurk: promote(9), cause(4), stimulate(3), regulate(2), induce(1), help in(1), affect(1), aid(1), spark(1), supplement(1).

Web: stimulate(403), regulate(222), promote(211), affect(105), control(67), cause(61), inhibit(31), influence(28), include(28), encourage(22), ...

0.68 "drug deaths" CAUSE_2

MTurk: be caused by(12), be due to(4), result from(3), occur because of(2), happen because of(2), occur from(1), emanate from(1), be from(1).

Web: be caused by(25), be(14), result from(9), involve(9), stand for(8), be fueled by(7), mention(6), occur without(6), be blamed on(4), be associated with(4), ...

0.68 "government land" HAVE_2

MTurk: be owned by(14), belong to(9), come from(1).

Web: be owned by(205), be acquired by(80), have(58), be(41), be taken by(39), be purchased by(34), be granted by(27), be leased from(26), be held by(23), be acquired from(21), ...

0.68 "picture album" FOR

MTurk: contain(18), hold(7), include(1), be filled with(1), have(1).

Web: contain(318), have(310), include(193), need(42), be(36), hold(34), feature(31), show(29), be filled with(28), display(19), ...

0.67 "shock treatment" USE

MTurk: use(10), involve(2), cause(2), include(1), induce(1), provide(1), invloves(1), be comprised of(1), administer(1), contain(1), ...

Web: include(13), use(13), involve(8), be(8), cause(5), produce(4), remove(4), be used during(3), induce(2), ensure(2), ...

0.66 "college town" HAVE_1

MTurk: contain(9), have(8), support(1), host(1), be attached to(1), accomodates(1), begin because of(1), nourish(1), sorround(1), encompass(1).

Web: have(94), be(22), house(10), dissect(8), attend(7), host(6), contain(5), support(5), accredit(4), center around(4), ...

0.63 "coke machine" FOR

MTurk: dispense(15), sell(3), vend(2), distribute(2), serve(1), give(1), contain(1), offer(1).

Web: sell(25), dispense(16), give(9), offer(7), draw(7), supply(6), have(4), be camouflaged as(4), serve(3), carry(3), ...

0.60 "winter season" BE

MTurk: include(4), be(4), occur in(2), be in(2), occur during(2), come in(2), describe(2), consist of(2), coincide with(1), feel like(1), ...

Web: be(128), occur in(20), be during(17), be in(15), follow(15), coincide with(14), be called(13), include(12), join(10), run from(9), ...

0.60 "bear country" HAVE_1

MTurk: have(8), contain(6), harbor(3), nourish(2), be inhabited by(2), feed(1), nurture(1), be populated with(1), love(1), be home to(1).

Web: have(33), do(15), protect(10), be(6), go(5), be screened in(3), deliver(3), sustain(3), see as(3), allow(3), ...

0.59 "lightning rod" FOR

MTurk: attract(14), conduct(3), channel(3), direct(2), arrest(2), draw(1), safeguard from(1), deflect(1), contain(1), invite(1), ...

Web: attract(17), be struck by(13), be melted by(9), convey(9), turn away(8), prevent(5), conduct(3), capture(3), ground(2), bear(2), ...

0.59 "designer creations" NOMINALIZATION:PATIENT

MTurk: be made by(14), come from(3), be created by(2), be produced by(2), be engineered by(1), be designed by(1), produce by(1).

Web: be made by(5), be produced by(4), be selected with(4), be created by(3), leave(3), require(3), come from(2), land(2), design(2), be done by(2), ...

0.59 "neighborhood bars" IN

MTurk: be in(8), be located in(4), be found in(2), be situated in(1), protect(1), exist within(1), cater to(1), encircle(1), reside near(1), reside in(1), ...

Web: be in(12), open in(8), ground in(5), typify(5), be located in(4), rock(4), pop up in(3), open up in(3), shutter in(3), give(3), ...

0.57 "rye whiskey" FROM

MTurk: be made from(14), be made of(4), come from(3), be made with(2), be distilled from(1), be compose of(1), be made using(1).

Web: be distilled from(4), be made from(3), call(1), contain(1), accent(1), have(1).

0.56 “marine life” IN

MTurk: be(7), exist in(2), characterize(1), be lived by(1), occur in(1), surprise(1), turn(1), refer to(1), emanate from(1), be led by(1), ...

Web: be(3), help(3), use(2), exist in(1), characterize(1), support(1), float in(1), showcases(1), end(1).

0.55 “sister node” BE

MTurk: be(3), accompany(2), act as(2), have(1), refer to(1), be related like(1), pertain to(1), be of(1), work with(1), be like(1), ...

Web: be(105), have(4), be contained in(4), be inserted as(3), be adjoined as(3).

0.55 “oil well” FOR

MTurk: contain(9), produce(6), pump(3), draw(2), spew(1), have(1), dispense(1), pump up(1), harness(1).

Web: produce(689), find(78), test(58), pump(44), be(35), flow(34), be drilled for(32), contain(23), yield(22), have(21), ...

0.55 “cane sugar” FROM

MTurk: come from(14), be made from(4), be processed from(2), be created from(2), be derived from(1), be made of(1), be produced from(1).

Web: be(18), come from(14), be made from(6), replace(6), be refined(5), predominate over(5), be in(4), be unlike(4), be derived from(3), compete with(3), ...

0.55 “country butter” FROM

MTurk: be made in(11), come from(8), be produced in(2), taste(1), be from(1), come out of(1), remind of(1).

Web: be produced in(12), come from(8), be made in(3), be preferred in(3), be imported into(3), be picked up around(2), be consumed in(2), be celebrated around(2), enter(1), leave(1), ...

0.54 “garter snake” BE

MTurk: resemble(8), look like(5), be called(2), like(1), be thin like(1), be named for(1), wear(1), be banded like(1), crawl through(1), have(1).

Web: resemble(5), be(5), include(4), look like(2), find(2), be triggered by(2), be confused with(2), eat(1), be construed as(1), sell(1).

0.54 “sports magazine” ABOUT

MTurk: feature(5), discuss(3), be about(3), cover(2), focus on(2), tell about(2), deal with(1), involve(1), write on(1), review(1), ...

Web: cover(95), feature(46), focus on(26), include(13), promote(12), deal with(11), depict(9), make(8), be(8), have(5), ...

0.54 “sob story” CAUSE$_1$

MTurk: make(5), cause(5), evoke(4), induce(3), elicit(3), produce(2), make for(1), provoke(1), cause to(1), be made of(1), ...

Web: make(15), produce(2), be interrupted by(1), turn into(1), come through(1).

0.54 “lion cub” BE

MTurk: come from(7), be(5), become(4), grow into(3), be birthed by(3), be born of(2), bear of(1), be born from(1), be sired by(1), descend from(1).

Web: become(6), be(4), grow into(3), blurt out(2), meet with(2), follow(2), belong in(2), have(2), make up(1), be threatened by(1), ...

0.53 “country visitors” FROM

MTurk: come from(6), go to(5), visit(4), be from(3), hail from(1), arrive from(1), visit from(1), come to(1), go in(1), originate in(1), ...

Web: come from(135), enter(133), visit(44), be in(42), leave(36), be from(26), come into(24), reside in(21), live in(18), stay in(18), ...

0.53 “mining engineer” FOR

MTurk: specialize in(4), study(3), know about(3), do(2), focus on(2), specialise in(1), understand(1), work with(1), know(1), work in(1), ...

Web: specialize in(15), work in(10), be(7), stay within(6), work for(6), do(5), work with(5), build(3), select(3), understand(2), ...

0.53 “vegetable soup” HAVE_1

MTurk: contain(8), be made from(5), be made of(3), come from(3), consist of(1), have(1), be composed of(1), have within(1), be comprised of(1), be made up of(1), ...

Web: contain(45), have(33), include(32), be made with(21), be(19), use(15), be filled with(10), be loaded with(8), puree(7), consist of(5), ...

0.53 “winter sports” IN

MTurk: be played in(11), occur in(5), happen in(3), be played during(2), happen during(2), be done in(1), utilize(1), be in(1), be enjoyed in(1), occur during(1).

Web: be played in(7), move from(6), be done in(4), enjoy in(3), be prepared by(3), be played during(2), be divided into(2), come with(2), dominate(2), play(2), ...

0.52 “star shape” BE

MTurk: resemble(12), look like(10), be like(1), recall(1).

Web: include(16), resemble(11), look like(8), be(6), represent(5), seep down from(5), blot out(4), create as(4), form(4), vary from(4), ...

0.52 “faculty decisions” NOMINALIZATION:PRODUCT

MTurk: be made by(23), affect(2), come down from(1), concern(1).

Web: affect(60), be made by(33), go against(9), maintain(7), allow(7), be criticized by(6), be(6), be supported by(5), restrict(5), rest with(5), ...

0.52 “frog man” BE

MTurk: resemble(6), look like(4), swim like(4), be like(2), be(2), emerge from(1), dive like(1), behave like(1), emulate(1), love(1), ...

Web: look like(24), be(12), find(12), resemble(11), hate(7), live on(7), blow up(7), swallow(7), find frog(7), write about(5), ...

0.51 "family problems" IN

MTurk: involve(4), occur in(2), plague(2), affect(2), be created by(1), include(1), be within(1), be suffered by(1), arise in(1), occur within(1), ...

Web: affect(318), plague(120), arise in(90), bring(90), confront(79), occur in(51), run in(47), involve(41), cause(39), threaten(33), ...

0.50 "flounder fish" BE

MTurk: be(8), look like(2), be sold as(1), be called(1), come from(1), appear(1), struggle like(1), be classified as(1), be seen in(1), love to(1), ...

Web: look like(9), be(4), resemble(4), include(2), be used include(2), substitute for(1), go with(1), be shaped like(1), follow(1), want(1).

0.50 "plastic toys" $MAKE_2$

MTurk: be made of(15), be made from(4), contain(2), consist of(1), be(1), be formulated of(1), be manufactured from(1).

Web: be(51), be made of(25), have(10), be made from(8), consist of(4), be made out of(3), heat(3), fire(3), be constructed of(2), be constructed from(2), ...

0.47 "food supplies" NOMINALIZATION:PRODUCT

MTurk: consist of(6), include(3), contain(3), be made up of(3), be(2), distribute(2), be made from(2), be comprised of(2), be made of(2), involve(1), ...

Web: include(58), be(34), make(17), consist of(11), provide(9), be used for(8), have(6), be derived from(6), mean that(5), flow around(5), ...

0.46 "census taker" NOMINALIZATION:AGENT

MTurk: take(4), count(2), collect data for(1), participate in(1), administer(1), canvas for(1), pertain to(1), complete(1), work on(1), document(1), ...

Web: take(2), mean that(2).

0.46 "fatigue headache" $CAUSE_2$

MTurk: be caused by(13), result from(4), come from(3), be brought on by(1), cause(1), occur do to(1), bring on by(1), come froom(1).

Web: be induced by(3), be caused by(2), result from(2), result in(2), be preceded by(2), be aggravated by(1), arise from(1).

0.46 "coal dust" FROM

MTurk: come from(15), be(2), be generated from(1), be attached to(1), be made from(1), be composed of(1), be spread by(1), emerge from(1), be made of(1), derive from(1), ...

Web: come from(5), be associated with(5), fuel(5), be generated from(4), be(4), contain(4), fall from(3), be caused from(2), feed(2), be correlated with(2), ...

0.46 "picture book" $HAVE_1$

MTurk: contain(15), have(3), be composed of(1), rely on(1), use(1), consist of(1), include(1), caintains(1), be made of(1).

Web: have(1146), contain(366), include(345), give(290), show(201), be(157), paint(119), present(68), use(67), provide(60), ...

0.45 "fish scales" $HAVE_2$

MTurk: be on(6), cover(5), be found on(4), grow on(3), come from(2), fall from(1), protect(1), weigh(1), shed from(1), be in(1), ...

Web: be found on(17), be embedded in(17), cover(14), weigh(12), sustain(9), make(7), attract(7), look like(5), allow(5), protect(4), ...

0.45 "sports activities" BE

MTurk: involve(8), include(3), be comprised of(2), be(2), revolve around(1), be composed of(1), center around(1), come from(1), depend on(1), work on(1), ...

Web: include(494), be(118), range from(70), involve(62), use(29), be considered(29), bring(16), organize(13), be classified as(13), be associated with(12), ...

0.44 "child actor" BE

MTurk: be(22), involve(1), portray(1), be classified as(1).

Web: play(113), be(65), have(32), portray(31), start as(18), work with(18), deal with(17), write(17), look like(15), make(13), ...

0.43 "suspense film" $CAUSE_1$

MTurk: contain(7), have(5), create(4), be(2), cause(2), build(1), include(1), provide(1), generate(1), be filled with(1), ...

Web: rely on(26), have(24), build(22), create(16), combine(14), use(9), be(8), lack(8), miss(8), be filled with(7), ...

0.42 "finger cymbals" USE

MTurk: be worn on(5), be played by(4), be played with(3), go on(2), use(2), attach to(2), be used by(2), wear on(1), involve(1), be(1), ...

Web: go on(1), be played with(1).

0.41 "cable network" $MAKE_2$

MTurk: consist of(3), run on(2), use(2), provide(2), carry(2), be on(1), run through(1), be connected by(1), be shown on(1), be carried on(1), ...

Web: use(154), include(52), be(31), consist of(25), provide(19), run over(18), employ(15), run(15), have(13), compete in(13), ...

0.41 "satellite nation" BE

MTurk: be(4), act like(3), consist of(2), have(1), be like(1), use(1), depend on like(1), be composed of(1), be made up of(1), look like(1), ...

Web: launch(23), be(20), put(7), use(6), have(5), exist(5), bring(5), be transmitted via(4), depend on(4), be governed as(3), ...

0.41 "child abuse" NOMINALIZATION:ACT

MTurk: happen to(6), affect(5), involve(2), concern(2), hurt(1), be perpetrated on(1), be directed at(1), target(1), occur with(1), occur to(1), ...

Web: affect(76), cause(30), place(30), be(29), involve(27), put(27), harm(26), drive(25), teach(23), be inflicted upon(20), ...

0.40 “cactus plant” BE

MTurk: be(12), look like(2), be composed of(1), derive from(1), relate to(1), bear(1), contain(1), be classified as(1), be categorized with(1), be related to(1), ...

Web: resemble(83), look like(77), be(38), include(13), grow like(7), be confused with(5), last(4), be among(4), associate with(3), range from(3), ...

0.39 “pine tree” BE

MTurk: have(4), smell like(3), be(3), be made of(3), be called(2), smell of(2), produce(2), be made from(1), belong to(1), be named(1), ...

Web: be(88), include(37), resemble(25), look like(18), surpass(10), be called(6), smell like(6), be planted(6), replace(6), crowd(6), ...

0.38 “mountain lodge” IN

MTurk: be located in(7), be in(7), be situated on(2), be on(1), be built on(1), be built in(1), be located on(1), be situated in(1), be up in(1), sit in(1).

Web: have(6), be located in(4), curse(4), be in(3), be nestled in(3), sit atop(3), operate on(3), overlook(3), be styled like(3), be tucked in(2), ...

0.38 “automobile plant” FOR

MTurk: manufacture(12), make(10), build(2), produce(1), assemble(1), create(1).

Web: produce(16), be utilized in(14), build(7), manufacture(7), assemble(5), include(5), make(3), turn out(2), be used in(2), process(2), ...

0.38 “mother church” BE

MTurk: act as(6), be(3), act like(3), resemble(2), be like(1), be originated with(1), orginates from(1), recruit(1), remind of(1), operate as(1).

Web: be(447), support(20), become(19), have(16), be brought up with(16), be represented as(11), help(10), be like(9), tell(8), find(7), ...

0.37 “policy matters” ABOUT

MTurk: involve(4), concern(4), relate to(2), pertain to(2), be related to(2), be of(1), have(1), regard(1), affect(1), be relative to(1), ...

Web: affect(134), involve(75), be(52), be covered by(44), have(33), include(17), depend upon(14), require(13), write(12), conflict with(11), ...

0.37 “home remedy” FROM

MTurk: be made at(9), come from(4), be used in(3), originate at(2), be done at(1), be used at(1), be found at(1), originate from(1), be made in(1), be provided at(1), ...

Web: be prepared at(16), be in(14), be tried at(11), be made at(10), be used in(6), be done at(6), be used at(4), build(4), be(4), be administered at(4), ...

0.37 “city planner” NOMINALIZATION:AGENT

MTurk: plan(9), design(4), work for(3), lay out(2), work with(2), study(1), come from(1), live in(1), supply(1), organize(1), ...

Web: work with(27), work for(24), serve on(22), work in(14), design(13), treat(12), build(11), work at(9), understand(7), deal with(7), ...

0.37 "enemy invasion" NOMINALIZATION:ACT

MTurk: come from(7), involve(6), be made by(4), consist of(2), be composed of(1), be initiated by(1), invade(1), be accomplished by(1), feature(1), be done by(1), ...

Web: be made by(4), make(4), be kept upbj(4), come from(3), be attempted by(3), drive(3), control(2), be aimed at(2), be visited upon(2), be held out by(1), ...

0.35 "queen bee" BE

MTurk: be(13), act as(2), grow into(1), look like(1), act like(1), rule as(1), come from(1).

Web: have(26), lose(26), be(19), surround(9), serve(9), feed(8), become(8), be with(7), be without(7), be separated from(7), ...

0.34 "tuition subsidies" NOMINALIZATION:PRODUCT

MTurk: pay for(5), fund(3), lower(1), reduce(1), mitigate(1), decrease(1), be funded by(1), cover(1), help(1), support(1), ...

Web: keep(7), pay for(4), enable(3), lower(2), approximate(2), include(2), reduce(1), make(1), pay(1).

0.34 "paper money" MAKE_2

MTurk: be made of(11), be made from(8), be(3), be printed on(2), be composed of(1), be made out of(1), be produced on(1).

Web: be(22), exist on(12), be spent on(11), be made from(6), exist in(5), give(5), be wrapped in(5), remain(5), pass along(5), say(5), ...

0.33 "extension ladder" HAVE_1

MTurk: have(11), utilize(2), provide(2), allow(1), form(1), include(1), contain(1), pull out(1), perform(1), supply(1), ...

Web: be(12), be used as(8), have(6), form(4), be generated by(4), be designed as(3), double as(3), be arranged in(3), be like(2), be utilized as(2), ...

0.33 "copper coins" MAKE_2

MTurk: be made of(11), be made from(8), look like(1), contain(1), be composed of(1), utilize(1), be made out of(1).

Web: be(36), be made of(14), be of(7), have(7), contain(4), be made from(3), look like(3), depress(3), be struck in(3), show(3), ...

0.32 "headache pills" FOR

MTurk: cure(11), relieve(4), treat(3), help with(1), help(1), be for(1), reduce(1), deaden(1), feed(1), negate(1), ...

Web: give(38), cause(19), cure(15), take away(9), develop(8), relieve(5), stop(5), rid of(4), be for(3), prevent(3), ...

0.31 "love song" ABOUT

MTurk: be about(12), speak of(3), talk about(2), be written about(1), describe(1), profess(1), speak about(1), express(1), sing about(1), concern(1), ...

Web: make(127), be(123), do(59), be about(58), celebrate(49), express(48), deal with(40), describe(33), speak of(32), go(32), ...

0.31 "oil imports" NOMINALIZATION:PRODUCT

MTurk: be(4), bring in(3), consist of(2), contain(2), involve(2), comprise(2), include(1), be of(1), supply(1), be composed of(1), ...

Web: include(23), consist of(5), be(5), constitute(5), result in(4), be of(3), be purchased through(3), exclude(3), offset(2), account for(2), ...

0.30 "vacuum cleaner" USE

MTurk: use(5), clean by(2), utilize(2), employ(1), be(1), do(1), suck into(1), suck by(1), pick up with(1), work on(1), ...

Web: be(44), invent(12), use(11), be like(7), produce(6), make(6), have(5), resemble(5), pick up than(5), be used as(4), ...

0.29 "worker teams" MAKE_2

MTurk: be comprised of(5), be made up of(4), be composed of(2), include(2), have(2), comprise(2), contain(1), use(1), bring(1), consist of(1), ...

Web: include(407), consist of(94), be(31), have(13), bring(10), provide(10), give(8), allow(8), incorporate(7), need(7), ...

0.28 "peanut butter" FROM

MTurk: be made from(10), come from(5), be made of(5), contain(3), be processed from(2), be made with(1), be composed of(1), be used with(1), be formulated from(1).

Web: contain(34), be(26), have(19), be made from(7), taste like(5), use(4), be than(4), resemble(3), list(3), produce(3), ...

0.28 "steam iron" USE

MTurk: use(7), produce(6), emit(3), utilize(3), make(1), incorporate(1), steam(1), puff(1), create(1), vent(1), ...

Web: generate(14), deliver(7), emit(4), produce(4), do(4), reduce(4), use(3), spray(3), operate on(3), discharge(3), ...

0.28 "history conference" ABOUT

MTurk: discuss(4), focus on(4), be about(4), deal with(3), concern(3), involve(2), study(1), be on(1), be related to(1), pertain to(1), ...

Web: have(36), focus on(22), explore(15), make(12), look at(11), go down in(8), go in(8), consider(8), deal with(6), focus upon(6), ...

0.28 "cover designs" NOMINALIZATION:PRODUCT

MTurk: appear on(4), be on(4), decorate(3), create(2), be for(2), be used for(1), serve(1), emerge from(1), be published on(1), be found in(1), ...

Web: allow(70), appear on(42), grace(33), have(16), become(16), make(15), feature(13), include(12), do(10), be used for(9), ...

0.27 "budget speech" ABOUT

MTurk: be about(8), discuss(4), concern(3), be on(1), describe(1), turn in(1), refer to(1), be made on(1), detail plan for(1), pertain to(1), ...

Web: affect(5), be on(3), be about(2), be taken in(2), be(2), be manifested through(1), be in(1), reflect(1), be delivered on(1).

0.27 “artifact descriptions” NOMINALIZATION:PRODUCT

MTurk: be about(5), describe(5), refer to(2), be of(2), define(2), revolve around(1), deal with(1), regard(1), about(1), tell about(1), . . .

Web: be(5), deploy(5), be about(4), relate(2), allow(2), handle(1), be based on(1).

0.27 “horse doctor” FOR

MTurk: treat(15), care for(3), tend to(2), heal(2), serve(1), cure(1), fix(1), practise on(1), work on(1).

Web: ride(9), be(9), see(8), travel by(8), treat(7), have(7), use(6), dislike(5), come by(4), amputate(4), . . .

0.27 “adventure story” ABOUT

MTurk: contain(6), describe(3), tell of(3), have(3), be about(3), involve(2), include(1), feature(1), be(1), tell(1), . . .

Web: follow(124), be(54), tell of(51), recount(49), relate(49), combine(37), chronicle(34), continue(31), tell(29), describe(27), . . .

0.27 “student committee” MAKE_2

MTurk: be comprised of(5), be made up of(5), consist of(3), contain(3), be made of(3), be composed of(2), involve(2), include(1), represent(1), be headed by(1).

Web: include(523), consist of(123), allow(100), be comprised of(76), meet with(65), be(64), select(64), advise(62), guide(60), be composed of(57), . . .

0.26 “teaching profession” BE

MTurk: involve(12), require(3), include(2), espouse(2), perform(1), enjoy(1), specialize in(1), revolve around(1), be devoted to(1), be about(1), . . .

Web: be(52), include(23), involve(12), have(12), be engaged in(10), enjoy(6), require(5), be liberalized by(5), recognize(5), occur in(3), . . .

0.26 “college employees” NOMINALIZATION:PATIENT

MTurk: work for(10), work at(8), be in(2), work in(2), be hired by(2), be enrolled in(1), go to(1), help out at(1).

Web: leave(95), be employed by(69), graduate from(63), attend(50), work for(32), serve(29), work at(26), have(23), be enrolled in(22), retire from(22), . . .

0.26 “bacon grease” FROM

MTurk: come from(17), be derived from(2), be made from(2), be rendered from(1), stick to(1), be exuded by(1), derive from(1), make(1), be produced by(1), be fried out of(1).

Web: be rendered from(4), run off(2), come from(1), sit on(1).

0.25 “nut bread” HAVE_1

MTurk: contain(11), be made with(2), consist of(2), be made of(2), be made using(2), include(1), be filled with(1), have(1), be composed of(1), be made from(1), . . .

Web: have(8), include(5), feature(4), be mixed with(4), be crammed with(3), load with(3), contain(2), be filled with(2), be stuffed with(2), be loaded with(2), . . .

0.24 "starvation diet" USE

MTurk: cause(3), mimic(2), require(2), lead to(2), come from(2), be like(2), induce(2), use(2), involve(1), be(1), ...

Web: be(14), seem like(8), involve(7), prevent(6), tend toward(5), require(4), mean(4), mimic(3), advocate(3), be near(3), ...

0.24 "book requests" NOMINALIZATION:PRODUCT

MTurk: ask for(6), be for(4), be made for(4), involve(2), concern(2), be made about(1), refer to(1), be related to(1), pertain to(1), want(1), ...

Web: be on(11), be for(4), publish(3), find(3), ask for(2), appear in(2), include(2), borrow(2), use(2), replicate(2), ...

0.23 "candy cigarette" $MAKE_2$

MTurk: be made of(13), taste like(4), be made from(4), make from(1), consist of(1), be(1), look like(1), be made with(1), resemble(1).

Web: taste like(3), have(2).

0.23 "cream sauce" $HAVE_1$

MTurk: contain(9), be made of(5), be made from(2), be like(1), use(1), be composed of(1), be shaped as(1), come from(1), be of(1), resemble(1), ...

Web: combine(19), have(17), be(14), include(8), contain(6), call for(6), be enriched with(5), taste like(5), need(5), be made from(4), ...

0.22 "stone tools" $MAKE_2$

MTurk: be made of(15), be made from(5), work with(1), shape(1), cut(1), carve away(1), work on(1).

Web: use in(52), cut(13), be made of(13), be used in(13), break(10), become(8), remove(8), be(7), move(5), bore into(5), ...

0.22 "hairpin turn" BE

MTurk: resemble(9), look like(7), mimic(2), be like(1), use(1), be sharp like(1), turn as(1), be curved like(1), be lik(1), remind of(1).

Web: promote(7), be(5), help for(3), resemble(2), look like(1), become(1).

0.22 "pressure cooker" USE

MTurk: use(9), cook with(6), utilize(2), cook under(2), produce(1), generate(1), work with(1), operate use(1), involve(1), cook by(1), ...

Web: operate at(9), maintain(7), use(4), prevent(4), control(4), be(4), generate(3), increase in(3), provide(3), increase(2), ...

0.22 "olive oil" FROM

MTurk: come from(11), be made from(4), be made of(3), be derived from(1), be pressed from(1), be(1), contain(1), be extracted from(1), be pressed out of(1), smell of(1).

Web: be(74), include(31), be extracted from(22), be sold as(21), resemble(17), taste like(16), be obtained from(14), be made from(13), have(12), be pressed from(10), ...

0.21 “morning prayers” IN

MTurk: be said in(8), be done in(2), be said(2), occur at(2), be made in(2), happen in(2), be recited in(1), be chanted in(1), be recited during(1), be meant for(1), ...

Web: be held(16), be read(13), be recited(10), have(10), be offered(10), be said(9), be said in(6), go up(6), mark(6), include(6), ...

0.21 “bronze statue” MAKE_2

MTurk: be made of(14), be cast from(2), be composed of(1), be(1), be made out of(1), be made from(1), be constructed from(1), be manufactured from(1), contain(1), be fabricated from(1), ...

Web: be cast in(52), be of(29), be(20), be made of(12), be in(9), be cast from(6), be cast of(3), become(3), inform(3), penetrate(3), ...

0.21 “soldier ant” BE

MTurk: act like(10), be(4), behave like(4), act as(2), work like(2), march on like(1), mimic(1), emulate(1), play(1), serve as(1), ...

Web: bring(5), be(4), be killed by(3), act as(2), be turned into(2), be called(1), approach(1), go as(1).

0.21 “ceiling price” BE

MTurk: hit(5), be(3), reach(2), form(2), go to(2), be at(2), constitute(1), serve as(1), be like(1), be terminated at(1), ...

Web: exceed(31), be(7), establish(7), set(7), be above(7), hit(6), act as(5), create(4), be below(4), go through(4), ...

0.20 “juvenile court” FOR

MTurk: try(8), handle(3), deal with(2), prosecute(2), involve(2), process(2), adjudicate(1), specialize in(1), adjucates for(1), govern(1), ...

Web: include(29), state that(29), be(25), place(18), commit(18), deal with(15), adjudicate(13), handle(13), sit as(12), confirm(9), ...

0.20 “sea breeze” FROM

MTurk: come from(21), blow off(1), come off(1), be near(1), come off of(1).

Web: blow from(72), come from(16), come off(15), blow towards(13), ruffle(13), make(12), cause(12), occur in(9), frilled(9), rise with(8), ...

0.20 “future shock” CAUSE_2

MTurk: involve(4), happen in(3), occur in(2), come from(2), be caused by(2), be relative to(1), remain(1), happen during(1), occur within(1), be linked to(1), ...

Web: occur in(10), affect(8), persist in(5), reverberate into(3), be(3), present(2), dissipate in(2), come in(1), be produced in(1), determine(1), ...

0.20 “cell division” NOMINALIZATION:ACT

MTurk: occur in(6), divide(3), happen in(3), happen to(2), affect(2), split(1), occur within(1), cut though(1), be made by(1), multiply(1), ...

Web: produce(173), result in(60), generate(51), occur in(38), divide(36), form(17), cause(14), turn(12), make(11), create(11), ...

0.20 “finger lakes” BE

MTurk: resemble(9), look like(6), be shaped like(4), flow in(2), contain(1).

Web: make up(15), look like(6), be(6), comprise(6), frame(4), stretch out like(3), precede(3), focus on(2), drain into(2), resemble(1), ...

0.20 “chocolate bar” MAKE$_2$

MTurk: be made of(11), be made from(9), contain(6), be composed of(2).

Web: be(103), contain(54), have(30), serve(26), taste like(13), combine(11), be made with(7), be sold in(6), be called(6), come in(6), ...

0.19 “phantom limb” BE

MTurk: be(7), act like(3), be like(2), feel like(2), appear(1), behave like(1), mimic(1), look like(1), exist as(1), resemble(1), ...

Web: produce(4), be(1), be known as(1).

0.19 “salt lake” HAVE$_1$

MTurk: contain(14), have(6), be made of(2), be made from(2), include(1), be composed of(1), be high in(1).

Web: be(71), fill(17), produce(15), have(12), contain(11), be located within(8), become(8), provide(6), afford(5), dissolve(5), ...

0.19 “gutter language” FROM

MTurk: come from(7), resemble(3), belong in(2), originate from(2), be heard in(1), be in(1), relate to(1), represent(1), pertain to(1), stem from(1), ...

Web: be of(3), be found in(2), savour of(2), settle in(2), embarrass(2), come from(1), belong in(1), rise from(1), escape from(1), be(1).

0.19 “store clothes” FROM

MTurk: be bought in(6), come from(4), be sold in(2), be found in(2), be bought from(2), be purchased in(2), be used in(1), be from(1), be located in(1), be bought(1), ...

Web: be in(46), be sold in(29), sell in(11), arrive in(10), come into(9), hit(9), be sold at(6), be sold by(6), go into(5), fit in(5), ...

0.19 “company assets” HAVE$_2$

MTurk: belong to(10), be owned by(6), be generated by(2), lie within(2), help(1), be(1), be held by(1), affect(1), be possessed by(1), come from(1).

Web: enable(173), allow(104), be owned by(77), require(62), affect(60), be used by(56), be adopted by(55), be held by(51), be(48), increase(39), ...

0.18 “birth pains” CAUSE$_2$

MTurk: be caused by(6), occur during(3), come from(2), accompany(2), happen during(2), precede(1), come before(1), occur with(1), coincide with(1), emit from(1), ...

Web: accompany(55), give(55), precede(16), follow(16), attend(14), come with(11), be suspended after(10), occur from(7), herald(5), arise from(4), ...

0.18 "city trainees" NOMINALIZATION:PATIENT

MTurk: work for(8), be trained by(4), be from(3), work in(2), live in(2), be trained in(2), be hired by(1), work at(1), be in(1), be taught by(1), ...

Web: work in(6), live outside(4), work at(3), complete(3), be hired by(2), be(2), join(1), be employed by(1).

0.18 "student problems" $HAVE_2$

MTurk: be had by(4), concern(3), involve(2), affect(2), belong to(2), occur with(1), originate from(1), be encountered by(1), frustrate(1), plague(1), ...

Web: require(507), affect(415), confront(209), prevent(163), allow(138), ask(116), interfere with(99), cause(95), engage(81), arise between(81), ...

0.17 "draft dodger" NOMINALIZATION:AGENT

MTurk: avoid(18), ignore(2), run away from(2), dodge(1), evade(1), hide from(1), run from(1), subvert(1).

Web: dodge(7), evade(6), protest(2), defy(2), avoid(1), prosecute(1).

0.17 "tire rim" $HAVE_2$

MTurk: surround(4), hold(3), be on(3), fit(2), cover(2), be of(2), protect(1), be in(1), involve(1), accept(1), ...

Web: have(27), be used with(27), carry(24), support(17), hold(15), cause(13), fit(9), need(9), retain(8), require(8), ...

0.16 "mountain range" $MAKE_2$

MTurk: be made of(5), contain(4), be composed of(2), consist of(2), be of(2), be(1), be created by(1), be comprised of(1), run through(1), be made up of(1).

Web: be(61), include(41), connect(16), have(15), contain(8), build with(8), be among(8), make(6), extend from(5), comprise(5), ...

0.16 "student inventions" NOMINALIZATION:PATIENT

MTurk: be made by(15), be created by(8), come from(2), be designed by(1).

Web: allow(18), enable(6), be made by(4), maximize(4), be developed by(4), be performed by(4), be discussed by(3), reduce(3), move(2), show(2), ...

0.16 "childhood dreams" IN

MTurk: occur in(5), occur during(3), come in(2), originate in(2), come from(1), have(1), start in(1), be dreamt in(1), revolve around(1), be had in(1), ...

Web: come from(13), begin in(9), haunt(8), start in(6), be(5), recur throughout(4), gild(3), begin from(3), be with(3), start at(3), ...

0.16 "hermit crab" BE

MTurk: act like(7), live like(4), resemble(4), be(3), hide like(1), be recluse like(1), mimic(1), emulate(1), be called(1), look like(1), ...

Web: be(3), become(3), be like(2), carry around(2), clasp(2), investigate(1), include(1), follow(1), ones(1).

0.15 “apple cake” HAVE$_1$

MTurk: be made from(6), be made with(5), contain(4), be made of(3), taste like(2), come from(2), use(1), originate from(1), be composed of(1), be baked with(1), . . .

Web: use(2), substitute(2), have(2), be made with(1), chop(1), consist of(1), include(1), dry(1), be(1), be shaped like(1), . . .

0.15 “movement schisms” IN

MTurk: divide(2), involve(2), split(1), occur to(1), emerge from(1), occur in(1), form within(1), be created by(1), be characterized by(1), seperate(1), . . .

Web: exist in(8), divide(7), beset(7), plague(6), wrack(4), rent(4), enervate(4), hobble(4), be mirrored in(3), splinter(3), . . .

0.15 “glass eye” MAKE$_2$

MTurk: be made of(18), be made from(4), be composed of(2), come from(2), be constructed of(1), be like(1), contain(1), be created from(1).

Web: need(35), be(24), wear(23), be of(18), look through(18), be hidden behind(17), require(16), be corrected with(16), have(12), be made of(11), . . .

0.15 “heat rash” CAUSE$_2$

MTurk: be caused by(24), come from(3), emerge from(1), result from(1), be cause by(1), be from(1).

Web: start as(3), look like(3), be aggravated by(3), occur in(2), be attended with(2), be associated with(2), be exacerbated by(2), be known(2), come from(1), be caused by(1), . . .

0.13 “sand dune” MAKE$_2$

MTurk: be made of(11), be formed of(2), contain(2), be composed of(1), be formed from(1), consist of(1), originate from(1), be made from(1), be comprised of(1), be made up of(1), . . .

Web: be(16), consist of(7), form in(6), move(5), be composed of(4), grow(4), accrete(3), have(3), be reported as(3), provide(3), . . .

0.13 “animal attack” NOMINALIZATION:ACT

MTurk: be made by(8), involve(4), come from(4), be caused by(3), be performed by(1), be brought on by(1), be perpetrated by(1), include(1), occur by(1), be done by(1), . . .

Web: make(5), appear in(4), leave(4), come from(2), provoke(2), deprive(2), be prevented by(2), cause(2), involve(1), be performed by(1), . . .

0.13 “plant food” FOR

MTurk: nourish(7), feed(7), be for(5), be used by(1), benefit(1), be given to(1), be made for(1), be given(1), be absorbed by(1), be supplied by(1), . . .

Web: come from(154), be derived from(74), kill(38), be(36), have(29), feed(26), operate(24), include(22), nourish(13), be used by(10), . . .

0.13 “milieu therapy” USE

MTurk: involve(3), use(2), describe(1), occur in(1), be based upon(1), change(1), consider(1), be based on(1), host(1), be performed in(1), . . .

Web: alter(3), modify(2), improve(2), involve(1), be like(1).

0.13 “immigrant minority” MAKE_2

MTurk: be comprised of(6), be made of(4), be composed of(3), consist of(3), be made up of(3), contain(2), be(1), form(1), have(1).

Web: be(68), be viewed as(5), compete with(4), be drawn from(4), include(3), compete against(3), be expelled as(3), descend from(3), be composed of(2), consist of(2), ...

0.11 “radio communication” USE

MTurk: come from(4), use(3), be transmitted over(1), be broadcast by(1), flow through(1), emerge from(1), transmit to(1), be made by(1), be by(1), emanate from(1), ...

Web: provide(29), include(28), be(27), be broadcast by(14), bring(13), be transmitted over(10), own(8), run over(7), use(6), involve(6), ...

0.11 “heart design” BE

MTurk: resemble(4), be shaped like(4), include(3), look like(2), incorporate(2), contain(2), involve(2), warm(1), be composed of(1), be made from(1), ...

Web: be(60), be at(59), capture(45), win(34), warm(28), include(21), lie at(19), form(18), touch(12), drive with(12), ...

0.10 “price dispute” ABOUT

MTurk: involve(8), be over(5), concern(3), be about(3), argue about(1), arise over(1), center on(1), question(1), argue over(1), arise from(1).

Web: regard(5), respect(3), translate into(3), be about(2), challenge(2), break out over(2), push(2), concern(1), exist over(1), increase(1), ...

0.10 “murder charge” BE

MTurk: involve(10), relate to(2), consist of(2), concern(2), be for(1), include(1), be labeled(1), indicate(1), come from(1), be of(1), ...

Web: include(228), attempt(34), be(18), commit(8), range from(8), instigate(8), be attempted(4), solicit for(3), be reduced from(3), involve(2), ...

0.10 “pet spray” FOR

MTurk: be used on(7), be for(2), go on(2), deter(1), be exuded by(1), be on(1), be made for(1), go onto(1), relate to(1), be applied to(1), ...

Web: interrupt(43), keep(18), deter(12), be used on(4), cleanse(3), condition(3), hydrate(2), harm(2), make(2), train(2), ...

0.09 “evening hours” IN

MTurk: occur in(9), be in(7), occur during(3), start in(1), include(1), be considered(1), fall during(1), indicate(1), pertain to(1), happen in(1), ...

Web: include(554), be(29), extend into(29), be in(11), be during(10), stretch into(9), run into(6), involve(6), require(6), start(5), ...

0.09 “subject deletion” NOMINALIZATION:ACT

MTurk: remove(6), involve(4), eliminate(4), include(3), delete(2), leave out(1), replace(1), erase(1), be of(1), concern(1).

Web: be(21), delete(10), be found in(2), be on(2), drop(2), limit(1), permit(1), affect(1), be about(1).

0.09 "people power" HAVE$_2$

MTurk: come from(7), belong to(3), emerge from(1), reside with(1), be made by(1), be within(1), be exercised by(1), be used by(1), reside in(1), drive(1), ...

Web: enable(108), allow(87), be(68), lead(66), make(64), bring(54), force(54), affect(48), attract(48), oppress(47), ...

0.09 "job tension" CAUSE$_2$

MTurk: be caused by(8), result from(4), come from(3), be created by(2), occur on(1), occur in(1), occur because of(1), occur during(1), occur at(1), be generated by(1), ...

Web: make(14), come from(5), cause(4), exist in(4), come with(2), harm(2), stir(2), go with(2), be(2), do(2), ...

0.09 "water drop" MAKE$_2$

MTurk: be made of(7), consist of(3), be composed of(3), be comprised of(2), be(1), contain(1), have(1), emerge from(1), swim in(1), hold(1), ...

Web: cause(23), be(13), fall into(13), make(10), bring back(8), suck in(8), contain(7), have(6), represent(6), grow in(6), ...

0.09 "police intervention" NOMINALIZATION:ACT

MTurk: involve(7), be done by(3), be made by(2), come from(2), be forced by(1), occur through(1), be initiated by(1), generate from(1), be held by(1), be formed by(1), ...

Web: bring(16), send(6), allow(6), involve(2), be implemented by(2), do iniohe(2), include(1), divert(1), assist(1), force(1).

0.08 "water wheel" USE

MTurk: move(4), use(3), turn use(2), spin(1), turn with(1), utilize(1), work in(1), be composed of(1), run on(1), tumble(1), ...

Web: receive(44), raise(39), lift(32), pump(28), be turned by(26), draw(24), spill(20), bring(13), be powered by(13), drive(11), ...

0.08 "wood shavings" FROM

MTurk: come from(7), be made of(7), be made from(4), contain(1), emerge from(1), be created by(1), be comprised of(1), be formed from(1), be formed by(1), be carved off(1), ...

Web: contain(5), be obtained from(2), make(1).

0.08 "kennel puppies" FROM

MTurk: live in(8), come from(5), be raised in(4), be born in(4), be found in(2), be bred in(2), be from(2), be kept in(1).

Web: pee in(110), be kept in(12), be raised in(11), be in(6), come from(4), love(4), live in(3), leave(3), be offered by(3), die in(2), ...

0.08 "water mark" CAUSE$_2$

MTurk: be made by(5), be caused by(3), indicate(2), look like(2), be made from(2), be created by(2), show(1), emboss(1), delineate(1), stain(1), ...

Web: be covered by(16), be(9), be removed with(8), be removed by(6), be inundated by(5), modify(5), show(4), have(4), be on(3), be washed out with(3), ...

0.08 "night flight" IN

MTurk: occur at(10), happen at(7), go at(2), fly during(1), fly at(1), begin at(1), fly in(1), happen by(1), originate in(1), come at(1), ...

Web: leave at(28), arrive at(11), be(11), be at(11), arrive(7), operate at(5), leave(5), launch at(5), fly at(4), occur at(3), ...

0.07 "parent organization" BE

MTurk: be composed of(4), be made up of(4), be for(2), be run by(2), consist of(2), include(1), involve(1), be(1), supervise like(1), be made for(1), ...

Web: help(366), work with(238), provide(207), assist(200), support(160), serve(154), be(136), represent(93), bring(87), educate(81), ...

0.07 "concrete desert" MAKE_2

MTurk: be made of(12), be covered in(1), be composed of(1), be comprised of(1), be in(1), include(1), be made up of(1), look like(1), be turned in(1), be fashioned from(1), ...

Web: be covered in(2), be(1).

0.07 "moth hole" CAUSE_2

MTurk: be made by(12), be caused by(3), be eaten by(3), be created by(2), come from(2), protect(1), be produced by(1), be eaten out by(1).

Web: let(3), be caused by(1), be(1).

0.07 "dream analysis" NOMINALIZATION:ACT

MTurk: explain(4), interpret(2), refer to(2), study(2), involve(2), include(1), investigate(1), describe(1), be applied to(1), be of(1), ...

Web: refer(7), have(6), be compared with(5), include(4), be given of(4), address(4), present(3), contain(2), embody(2), bring(2), ...

0.06 "office friendships" IN

MTurk: start in(2), begin in(2), occur in(2), be in(2), be formed in(2), form in(2), begin at(2), be made in(1), originate in(1), flourish in(1), ...

Web: improve by(4), extend outside(3), elevate(2), be cemented by(2), extend beyond(2), make(2), share(2), start in(1), be made in(1), transcend(1), ...

0.05 "birth control" NOMINALIZATION:ACT

MTurk: prevent(9), regulate(4), stop(3), limit(2), control(1), prohibit(1), emerge from(1), inhibit(1), impede(1), involve(1), ...

Web: give(92), acne(10), have(8), be held after(7), prevent(5), occur after(5), result in(4), be(4), rely on(4), make for(3), ...

0.05 "nose drops" FOR

MTurk: be used in(4), go in(4), come from(3), be placed in(2), treat(2), go into(2), be for(1), drain from(1), be inserted into(1), drop into(1), ...

Web: run down(6), be suspended from(3), pay thru(3), be inhaled through(3), hang from(2), fell from(2), go down(2), get into(2), be used in(1), hang at(1), ...

0.05 "student critiques" NOMINALIZATION:PRODUCT

MTurk: be made by(5), be written by(4), be created by(2), be performed by(2), originate from(1), be given to(1), criticise(1), insult(1), emaqnate from(1), be generated of(1), . . .

Web: allow(14), help(9), engage(4), be sent between(4), require(4), encourage(3), present(3), provide(3), motivate(3), be written by(2), . . .

0.05 "air pressure" CAUSE$_2$

MTurk: come from(6), be caused by(5), occur in(1), be exerted by(1), use(1), force(1), be measured in(1), effect(1), be made by(1), be induced by(1), . . .

Web: draw(109), cause(97), force(67), pull(41), prevent(34), be(27), drive(23), suck(19), contaminate(18), push(15), . . .

0.05 "student decision" NOMINALIZATION:ACT

MTurk: be made by(20), face(1), involve(1), come from(1), generate from(1), be taken by(1).

Web: affect(634), allow(80), support(42), benefit(41), be made by(31), require(29), involve(21), enable(21), impact(21), limit(18), . . .

0.04 "lace handkerchief" HAVE$_1$

MTurk: be made of(16), be made from(4), be trimmed with(1), be made out of(1), be comprised of(1), be trimmed in(1), come from(1), be decorated with(1), have(1).

Web: be trimmed with(1), feature(1).

0.04 "surface tension" IN

MTurk: be on(5), occur on(3), occur at(2), reside on(2), appear on(1), disrupt(1), exist on(1), affect(1), be found on(1), be experienced on(1), . . .

Web: lie beneath(69), simmer beneath(44), lurk beneath(30), lay beneath(23), simmer under(22), cause(22), be under(20), be below(18), simmer below(17), clean(16), . . .

0.04 "abortion problem" ABOUT

MTurk: involve(6), be related to(3), concern(3), occur with(2), come from(2), occur in(2), be caused by(1), surround(1), originate from(1), refer to(1), . . .

Web: cause(17), include(13), follow(9), be associated with(7), arise from(5), require(5), be caused by(4), occur during(4), result from(4), get(3), . . .

0.04 "faith cure" USE

MTurk: rely on(5), be caused by(3), involve(2), be done by(2), be related to(2), come from(1), require(1), rely upon(1), occur through(1), be ascribed to(1), . . .

Web: be wrought by(7), rest in(4), be activated by(3), be obtained by(3), establish(3), ignore(3), be performed by(3), require(2), come from(1), be produced through(1), . . .

0.04 "bull ring" FOR

MTurk: contain(8), house(2), be worn by(2), be in(1), be located in(1), be on(1), be for(1), feature(1), fence in(1), hold(1), . . .

Web: resemble(7), be(7), be in(5), grip(5), look like(5), face(4), form(4), comprise(3), produce(2), belong on(2), . . .

0.03 “cash basis” BE

MTurk: use(3), be based on(3), involve(3), come from(2), include(1), be backed by(1), depend on(1), determine(1), rely on(1), consist of(1), ...

Web: recognise(62), be(15), include(13), recognize(12), generate(11), measure(9), mean that(7), affect(7), charge(4), be in(4), ...

0.03 “oil crisis” ABOUT

MTurk: involve(7), be caused by(6), be about(2), depend upon(1), relate to(1), be precipitated by(1), deal with(1), be due to(1), be related to(1), be from(1), ...

Web: bear upon(12), be known as(11), follow(10), arise from(7), originate in(7), result from(6), impact(5), send(5), push(4), be(4), ...

0.03 “summer months” IN

MTurk: be in(7), occur in(5), occur during(3), happen in(2), begin in(1), appear in(1), come in(1), be during(1), contain(1), be included in(1), ...

Web: be(47), follow(24), include(17), end(8), start(7), make up(7), constitute(7), masquerade as(6), be employed during(5), endeth(5), ...

0.03 “machine translation” USE

MTurk: be done by(10), be performed by(4), be made by(3), come from(3), be generated by(1), generate from(1), emanate from(1), pertain to(1), be developed by(1), be specific for(1).

Web: use(9), make(8), have(6), allow(5), be(5), be called(4), take(4), run(4), be generated by(3), be known by(3), ...

0.03 “smoke signals” USE

MTurk: be made of(9), be made from(5), use(3), emerge from(2), be(1), be caused by(1), consist of(1), be comprised of(1), result in(1), signal(1), ...

Web: provide(5), raise(5), be(4), contain(4), be corrected by(4), be altered by(3), distinguish between(2), emerge as(2), be obscured by(2), include(1), ...

0.03 “land reclamation” NOMINALIZATION:ACT

MTurk: involve(6), reclaim(2), provide(1), be done to(1), get back(1), retrieve(1), emerge from(1), take back(1), free up(1), recover(1), ...

Web: require(13), withdraw(8), prevent(5), own(5), provide(4), merge in(4), return(4), make(3), manage(3), occur on(2), ...

0.03 “pole height” HAVE$_2$

MTurk: measure(3), reach(2), be set for(1), be reached by(1), be ascribed to(1), include(1), limit(1), come from(1), be measured on(1), determine(1), ...

Web: allow(6), cause(3), vary from(2), be set for(1).

0.02 “energy emergency” ABOUT

MTurk: involve(7), be due to(3), concern(3), relate to(2), be from(2), require(1), impact(1), restrict(1), refer to(1), pertain to(1), ...

Web: demand(16), call(7), require(3), give(3), deplete(2), tax(2), call for(2), affect(2), bed(1), allow(1).

0.01 "tax law" ABOUT

MTurk: be about(6), involve(5), govern(2), deal with(2), pertain to(2), be related to(2), justify(1), define(1), refer to(1), relate to(1), ...

Web: impose(476), reduce(94), raise(88), pay(62), levy(61), increase(41), place(31), require(30), establish(30), allow(30), ...

0.01 "party members" $HAVE_2$

MTurk: belong to(10), form(2), attend(2), go to(2), join(1), be in(1), participate in(1), come from(1), be of(1), make up(1), ...

Web: be(3555), leave(190), represent(186), join(165), attend(74), contract(72), become(62), switch(48), quit(35), support(34), ...

0.01 "hydrogen bomb" USE

MTurk: use(6), be made from(4), be made of(4), be composed of(2), utilize(2), contain(2), ignite(1), need(1), be fueled by(1), consume(1), ...

Web: squeeze(5), fuse(4), detonate(4), develop(3), combine(3), run on(2), be known as(2), make(2), be(2), oppose(2), ...

0.01 "voice vote" USE

MTurk: be made by(7), be given by(2), use(2), be done by(2), be taken by(1), count(1), occur through(1), determine by(1), be expressed by(1), be conducted by(1), ...

Web: be(13), give(9), represent(9), have(9), be rejected by(5), be by(4), muffle(3), carry by(3), provide(3), pass by(3), ...

0.01 "color television" $HAVE_1$

MTurk: display in(5), be in(4), display(3), show(2), contain(2), project in(2), show in(2), use(1), transmit in(1), include(1), ...

Web: have(10), describe(4), be(4), reflect(3), invent(2), use(1), render(1), flash(1), create(1), match(1), ...

0.01 "student power" $HAVE_2$

MTurk: come from(5), be possessed by(3), be held by(2), be exercised by(2), involve(2), inspire(1), be wielded by(1), belong to(1), be yielded by(1), derive from(1), ...

Web: enable(34), be(11), draw(10), keep(10), admit(8), exist between(8), take(8), arrest(7), make(7), allow(7), ...

0.01 "family antiques" $HAVE_2$

MTurk: belong to(9), be owned by(6), be kept in(1), come from(1), pass through(1), remain with(1), be linked to(1), be passed down within(1), be kept by(1), survive(1), ...

Web: be in(77), be passed down through(6), be kept in(4), be handed down through(4), belong in(4), come from(3), be honored in(2), be compiled since(1), be passed down(1), come into(1), ...

0.01 "sugar cube" $MAKE_2$

MTurk: be made of(14), be made from(5), consist of(3), be formed from(2), contain(1), be composed of(1), be comprised of(1), be all(1), store(1), be formed of(1).

Web: be dusted with(8), compose of(8), drip on(2), be dusted in(2), contain(1), be steeped in(1), look like(1), be covered in(1).

0.00 “plum wine” FROM

MTurk: be made from(15), come from(5), be made of(2), be made with(2), contain(1), taste like(1), be produced from(1).

Web: have(14), offer(9), taste of(9), show(6), take(6), evoke(5), be(4), be packed with(4), resemble(3), be loaded with(3), . . .

0.00 “abortion vote” ABOUT

MTurk: be about(4), support(3), involve(3), concern(3), decide on(1), prohibit(1), approve(1), be for(1), be against(1), be taken regarding(1), . . .

Web: legalize(6), further(4), outlaw(4), criminalize(2), allow(2), permit(1).

0.00 “adolescent turmoil” IN

MTurk: be experienced by(5), occur in(5), affect(4), happen to(3), be caused by(2), involve(2), effect(1), bother(1), occur to(1), emerge from(1), . . .

Web: accompany(2), surround(1).

0.00 “alligator leather” FROM

MTurk: be made from(11), come from(6), be made of(2), be on(1), belong to(1), be obtained from(1), be(1), be skin from(1), be from(1), be born of(1), . . .

Web: look like(2), include(1), resemble(1).

0.00 “apple core” HAVE$_2$

MTurk: come from(10), be in(3), be from(2), be located in(1), be contained within(1), be found in(1), be inside of(1), center(1), be inside(1), be within(1), . . .

Web: win(2), be used in(2), be removed from(2), show up in(1), be(1), run(1).

0.00 “arms budget” FOR

MTurk: be for(5), pay for(3), include(2), concern(2), be earmarked for(1), be meant for(1), be used for(1), be allocated to(1), relate to(1), consider(1), . . .

Web: be(4), play(3), cost(2), back(1), give(1), permit(1), go for(1).

0.00 “basketball season” FOR

MTurk: involve(4), feature(4), include(3), contain(3), be for(2), have(2), demonstrate(1), be devoted to(1), welcome(1), focus on(1), . . .

Web: be(6), be followed by(3), play(3), lift(2), care about(1), mean(1), come after(1).

0.00 “beard trim” NOMINALIZATION:PRODUCT

MTurk: shorten(3), be done to(2), cut(2), happen to(2), involve(2), affect(2), be given to(1), trim(1), be for(1), cut back(1), . . .

Web: .

0.00 “bird reproduction” NOMINALIZATION:ACT

MTurk: involve(5), resemble(2), reproduce(2), occur with(1), depict(1), produce(1), occur in(1), be made by(1), be created by(1), be specific to(1), . . .

Web: be imposed on(1), be adorned with(1), show(1).

0.00 “blanket excuse” BE

MTurk: cover like(8), act like(3), be like(2), be used as(1), behave like(1), serve to(1), be analogous to(1), be(1), be given as(1).

Web: .

0.00 “canine companion” BE

MTurk: be(20), derive from(1), assist as(1), descend from(1), be of(1).

Web: cheer on(3), play with(2).

0.00 “cigarette war” ABOUT

MTurk: involve(10), revolve around(2), be over(2), be about(2), ban(1), be fought because of(1), be fought over(1), be against(1), concern(1), oppose(1).

Web: democratise(1), give(1), get(1).

0.00 “coffee nerves” CAUSE$_2$

MTurk: be caused by(9), come from(5), be from(2), be caused from(1), run on(1), emanate(1), be affected by(1), be blamed on(1), recognize(1), be generated by(1), ...

Web: .

0.00 “communist tenet” IN

MTurk: be held by(5), be believed by(3), support(2), be followed by(1), be believed in by(1), belong to(1), be made by(1), be created by(1), be defined by(1), come from(1), ...

Web: .

0.00 “coriander curry” HAVE$_1$

MTurk: contain(8), be made from(5), come from(3), be garnished with(2), have(2), be flavored with(1), produce(1), feature(1), look like(1), be cooked with(1).

Web: roast(1).

0.00 “daisy chains” MAKE$_2$

MTurk: be made of(11), be made from(6), look like(2), be made up of(1), contain(1), cut(1), resemble(1), link(1), emulate(1), be created from(1), ...

Web: be(4), be stabilised by(4).

0.00 “disaster flick” ABOUT

MTurk: be about(4), involve(3), describe(2), show(2), portray(2), be written about(1), be created by(1), center around(1), be based on(1), simulate(1), ...

Web: be saved from(2), merge(2).

0.00 “enemy strength” HAVE$_2$

MTurk: be possessed by(6), belong to(4), be attributed to(1), be controlled by(1), be held by(1), arise out of(1), come from(1), characterize(1), be of(1), evaluate(1), ...

Web: destroy(9), bring(9), be turned against(6), deter(6), crush(6), be(4), bear down(4), overthrow(3), break(3), be used against(3), ...

0.00 “factory rejects” NOMINALIZATION:PATIENT

MTurk: come from(11), be made by(2), be from(2), appear in(1), be discarded by(1), be dismissed by(1), be denied by(1), be refused from(1), plague(1), be found in(1), ...

Web: .

0.00 “financing dilemma” ABOUT

MTurk: involve(6), be caused by(3), concern(3), be about(3), come from(2), cause(2), emerge from(1), occur in(1), relate to(1), complicate(1), ...

Web: .

0.00 “government employment” IN

MTurk: be provided by(4), involve(3), be sponsored by(2), be made by(2), be with(2), serve(1), be by(1), be within(1), be given by(1), happen for(1), ...

Web: include(12), be adopted by(12), be generated by(6), be registered with(4), be promoted by(4), ask(4), be(4), be slowed down by(4), reduce(4), be endorsed by(4), ...

0.00 “handlebar mustache” BE

MTurk: look like(11), resemble(10), be like(1), appear like(1), recall(1).

Web: twist into(1), hang in(1).

0.00 “heart massage” NOMINALIZATION:ACT

MTurk: stimulate(5), help(3), be performed on(2), soothe(2), restart(2), be transmitted by(1), concentrate on(1), be applied to(1), revive(1), vibrate(1), ...

Web: connect(8), nourish(5), bruise(3), reach(2), get into(2), slow(2), win(1), delight(1), work(1), be affirmed in(1).

0.00 “infant colonies” BE

MTurk: be composed of(3), contain(3), resemble(2), be like(2), be(2), behave like(1), be comprised of(1), be populated by(1), be made up of(1), be still(1), ...

Web: affect(1), have(1).

0.00 “jungle exploration” NOMINALIZATION:ACT

MTurk: occur in(5), involve(4), go into(4), be in(3), happen in(2), emerge from(1), go through(1), map(1), look at(1), get into(1), ...

Web: figure(1), ascend from(1).

0.00 “language riots” ABOUT

MTurk: involve(5), be caused by(4), concern(2), be about(2), be incited because of(1), be caused for(1), be inspired by(1), occur because of(1), happen because of(1), stem from(1), ...

Web: follow(2).

0.00 “laugh wrinkles” CAUSE_2

MTurk: be caused by(12), come from(8), result from(2), show as(1), be gotten from(1), be created by(1), occur from(1), appear with(1).

Web: add(1), make(1).

0.00 “lemon peel” HAVE$_2$

MTurk: come from(17), be found on(2), cover(2), come off(2), be on(1), be made from(1), protect(1), surround(1).

Web: .

0.00 “liquor orders” NOMINALIZATION:PRODUCT

MTurk: request(7), contain(6), be for(3), include(3), be made for(1), supply(1), consist of(1), involve(1), be made on(1), demand for(1), ...

Web: indulge in(3), intoxicate(2), permit(2), require intoxicating(1).

0.00 “midnight snack” IN

MTurk: be eaten at(15), occur at(3), be consumed at(2), happen at(2), be eaten around(1), occur around(1), around(1), be had at(1), eat in(1), go into(1), ...

Web: .

0.00 “morphology lecture” ABOUT

MTurk: be about(9), describe(4), discuss(2), include(2), explain(2), deal with(1), address(1), teach(1), teach about(1), cover(1), ...

Web: .

0.00 “onion tears” CAUSE$_2$

MTurk: be caused by(14), come from(3), be generated by(1), be created by(1), arise due to(1), be induced by(1), be provoked by(1), react to(1), flow with(1).

Web: come with(6), chop(4), live in(3).

0.00 “pedal extremeties” BE

MTurk: be(6), be used to(2), resemble(2), use(2), occur on(1), move(1), operate(1), include(1), be located on(1), look like(1), ...

Web: .

0.00 “pet theory” BE

MTurk: involve(5), be about(3), pertain to(2), apply to(1), be treated like(1), refer to(1), relate to(1), resemble(1), be favored like(1), be like(1), ...

Web: satisfy(4), line up with(3), revolve around(1), challenge(1).

0.00 “pork suet” FROM

MTurk: be made from(10), be made of(5), come from(4), be rendered from(2), be trimmed from(1), use(1), contain(1), be from(1), make of(1).

Web: .

0.00 “pot high” CAUSE$_2$

MTurk: come from(9), be caused by(3), result from(3), be derived from(1), utilize(1), be created by(1), derive from(1), be induced by(1), be from(1), be created from(1), ...

Web: extend over(3), take(2), score(2), lift(1), be kept in(1), make(1), melt(1), be planted in(1), keep(1), require(1).

0.00 “rice paper” FROM

MTurk: be made of(12), be made from(4), make from(1), be made out of(1), be composed of(1), in make from(1), be made up of(1), contain(1), make of(1), come from(1), ...

Web: be(7), domesticate(4), range from(2), affirm(2), have(2), be eaten with(1), be weighted down with(1), be called(1), precoated(1).

0.00 “ship landing” NOMINALIZATION:ACT

MTurk: be used by(4), be for(3), be made by(3), occur with(1), be made for(1), be meant for(1), assist(1), berth(1), come from(1), be provided for(1), ...

Web: be made in(3), bring(2), cause(2), look like(2), break(2), maneuver(2), allow(2), be with(1), damage(1), keep(1).

0.00 “staff attempts” NOMINALIZATION:ACT

MTurk: be made by(17), be done by(3), come from(3), originate from(1), be promulgated by(1), concern(1).

Web: upset(2), be made for(1).

0.00 “steel helmet” MAKE_2

MTurk: be made of(15), be made from(6), be created of(1), be made out of(1), contain(1), come from(1).

Web: be of(19), sparkle like(2), go under(2), be like(1), replace(1), fit inside(1).

0.00 “summer travels” IN

MTurk: occur during(10), happen in(5), occur in(3), be in(2), enhance(1), occur(1), be made in(1), involve(1), begin during(1), go into(1), ...

Web: begin in(3), take(3), increase in(3), be executed during(2), work for(2), be presented in(2), begin(2), be encountered during(2), run from(1), be embarked on(1), ...

0.00 “tape measure” BE

MTurk: use(4), be made of(4), be made from(2), be on(1), be in(1), be made by(1), be created out of(1), look like(1), be formed by(1), contain(1), ...

Web: reduce(22), cut(18), have(9), prevent(8), make(7), eliminate(5), protect(4), include(4), retract(4), involve(3), ...

0.00 “testtube baby” FROM

MTurk: be conceived in(12), be made in(4), be created in(4), come from(4), be born from(1), be fertilized in(1), be created with(1).

Web: .

0.00 “tobacco ash” FROM

MTurk: come from(15), emerge from(2), derive from(2), stick to(1), be made from(1), result from(1), be(1), be from(1), be created from(1), be produced from(1).

Web: fall from(5), be mixed with(2).

0.00 "vapor lock" CAUSE_2

MTurk: be caused by(7), contain(4), result from(3), control(2), be created by(1), form(1), seal in(1), come from(1), develop in(1).

Web: .

0.00 "warrior castle" MAKE_2

MTurk: house(6), belong to(4), contain(4), be occupied by(2), be inhabited by(2), hold(1), be ruled by(1), shelter(1), castle(1), have(1), ...

Web: be thronged with(3), be defended by(3), produce(1), allow(1).

0.00 "wastebasket category" BE

MTurk: belong in(4), include(3), go in(2), describe(1), be named(1), deal with(1), contain(1), relegate to(1), be named after(1), use(1), ...

Web: .

0.00 "weekend boredom" IN

MTurk: occur on(10), happen during(4), occur during(3), happen on(3), emerge at(1), be produced by(1), mar(1), characterize(1), experience in(1), come on(1).

Web: plague(2).

E.3 Comparison by Class: Using All Human-Proposed Verbs

0.89 CAUSE$_1$

Humans: cause(128), create(41), produce(21), make(20), promote(18), generate(18), result in(17), lead to(17), induce(15), contain(13), encourage(10), spread(9), stimulate(9), have(7), carry(7), involve(7), bring about(6), provoke(5), give(5), affect(4), evoke(4), be(4), inflict(4), instigate(4), elicit(4), influence(3), help(3), include(3), increase(3), trigger(3), be related to(3), infect with(3), engender(3), relate to(3), inspire(3), cause to(3), precipitate(2), regulate(2), enhance(2), initiate(2), enable(2), start(2), accelerate(2), allow(2), utilize(2), be marked by(2), foster(2), be made of(2), form(2), force(2), occur by(2), come from(2), supply(2), result(2), contribute to(2), fuel(1), speed(1), propagate(1), center around(1), help in(1), offer(1), control(1), be implicated in(1), support(1), pass(1), be filled with(1), build(1), be characterized by(1), turn into(1), build up(1), feed(1), provide(1), breed(1), be associated with(1), help with(1), transmit(1), use(1), look like(1), be condusive to(1), impart(1), exacerbate(1), make for(1), happen by(1), fish for(1), acclaim from(1), descend in(1), draw(1), emerge from(1), be comprised of(1), embody(1), fascilitates(1), appear(1), begin(1), be conceived by(1), irritate(1), project(1), incubate into(1), occasion(1), be cloaked in(1), pass along(1), appeal to(1), exude(1), bring on(1), be found in(1), be necessary for(1), sicken with(1), nourish(1), illicits(1), intensify be(1), intensify(1), deal with(1), be conducive to(1), deserve(1), aid(1), effect(1), s purpose be(1), invite(1), happen from(1), expect(1), be produced by(1), need(1), forbodes(1), be correlated with(1), be because of(1), manufacture(1), be steeped in(1), bring to(1), pertain to(1), showcases(1), be caused by(1), transfer(1), provide for(1), inswtigates(1), incite(1), perpetuate(1), preclude(1), compel(1), be indicated by(1), recruit(1), spark(1), supplement(1), quicken(1), percipitates(1), implement(1), disseminate(1), accompany(1), feature(1), be written for(1), shrades(1), be about(1).

Web: cause(1949), stimulate(403), regulate(222), promote(211), produce(109), affect(105), control(67), be(52), include(48), spread(44), have(42), inhibit(31), influence(29), rely on(26), create(25), induce(25), encourage(24), carry(24), give(24), build(22), trigger(22), make(21), accelerate(17), differentiate(17), stunt(16), combine(14), result in(13), spur(12), modulate(12), bring(11), govern(11), be required for(10), be like(10), limit(10), propagate(9), use(9), support(9), be associated with(9), prevent(9), determine(9), stop(9), help with(8), suppress(8), lack(8), work with(8), be involved in(8), choke as(8), miss(8), be filled with(7), fuel(7), offer(7), transmit(7), mimic(7), be needed for(7), drive(7), ensure(7), display(6), slow(6), be trapped in(6), act as(6), work as(6), become(6), initiate(5), generate(5), increase(5), modify(5), maintain(5), blend(5), be for(5), help(4), provide(4), contain(4), be implicated in(4), bring about(4), develop from(4), twist around(4), activate(4), curb(4), be used in(4), pass on(4), produce that(4), reduce(4), impair(4), tamponades(4), be confirmed by(4), capture(4), replace(4), deliver(4), build on(4), imply(4), stretch(4), allow for(4), suggest that(4), be characterized by(3), start(3), utilize(3), provoke(3), help in(3), enable(3), balance(3), improve(3), be mediated by(3), call(3), breast(3), isn(3), mine(3), contract(3), add(3), build in(3), do(3), cause that(3), sow(3), be concealed by(3), facilitate(3), understand(3), ignite(3), range from(3), be made from(3), spread like(3), depend on(3), be unlike(3), feed off(3), speed(2), pass(2), feed(2), involve(2), build up(2), be unleashed after(2), be focused on(2), retard(2), rack up(2), plod along with(2), result from(2), advance(2), resemble(2), shoot(2), be called(2), exist during(2), veer towards(2), aid in(2), appear as(2), express(2), start from(2), mix(2), pile on(2), be released include(2), arrest(2), livestock(2), disappear as(2), break(2), ...

0.86 MAKE$_1$

Humans: produce(46), make(44), sing(22), play(19), create(18), have(9), contain(9), emit(6), manufacture(6), exude(6), generate(6), provide(5), store(5), warble(4), chirp(4), secrete(4), supply(3), ooze(3), give(3), eat(2), deliver(2), be known for(2), give off(2), sound like(2), leak(1), run with(1), bleed(1), emanate(1), release(1), gather(1), abound in(1), include(1), drip(1), repeat(1), perform(1), be identified by(1), burst into(1), look like(1), amplify(1), utter(1), feature(1), be harvested for(1), drain out(1), eject(1), be famous for(1), issue(1), tweet like(1), excrete(1), vocalize(1), bleed out(1), bleed into(1), chant(1), croon(1), formulate(1), flow with(1), be praised for(1), maunfactures(1), cultivate(1), employ(1), communicate with(1), be productive of(1), be streaked with(1), ooze out(1), grow(1), leach out(1), communicate in(1), erupt in(1), synthesize(1), be made for(1), expel(1), chortle(1), give out(1), tweet(1), turn out(1), love(1), create sweet(1), make useable(1), be filled with(1).

Web: make(333), produce(295), sing(264), have(196), play(105), gather(104), learn(80), collect(57), force(52), be(45), hear(41), suck(40), give(35), drip(33), exude(33), provide(30), carry(26), get(26), bring(24), contain(20), ooze(20), draw(20), lose(20), supply(19), extract(19), cause(18), continue in(18), yield(18), create(17), know(17), secrete(16), drop(16), drink(16), allow(15), accumulate(13), be tapped for(12), bleed(11), raid(11), develop(11), eat(10), store(10), be kept for(10), be in(10), mimic(10), do(10), acquire(10), be attracted by(10), leak(9), look like(9), share(9), change(9), vary(9), look for(9), find(9), seek(9), turn(9), rely on(9), break into(8), access(8), resemble(8), consume(8), close(8), include(7), emit(7), prevent(7), add(7), exhaust(7), copy(7), feed on(7), burst with(7), sting(7), say(7), deliver(6), feature(6), live on(6), be without(6), leave(6), be known as(6), forget(6), offer(6), complete(6), feature in(6), amplify(5), flow(5), handle(5), master(5), prepare(5), modify(5), move(5), fill(5), simulate(5), steal(5), defend(5), fly like(5), send(5), sip(5), take(5), weep(5), bring in(5), hold(5), degrade(5), hive(5), warble(4), generate(4), be harvested for(4), utter(4), transport(4), lack(4), help(4), blast(4), taperecorded(4), call(4), feed(4), be called(4), send out(4), promise(4), wanteth(4), run on(4), reach(4), whisper(4), pick up(4), reproduce(4), store up(4), stream(4), be checked for(4), trill(4), roam for(4), use be(4), be cut off that(4), knead up(4), be known for(3), spread from(3), twitter(3), taste(3), come with(3), begin(3), lay up(3), improve(3), be on(3), represent(3), carry over into(3), combine(3), deposit(3), instill(3), use(3), discriminate between(3), possess(3), obtain(3), want(3), enjoy(3), ...

0.75 BE

Humans: be(272), look like(161), resemble(159), involve(68), act like(57), include(50), be like(39), come from(38), contain(33), be shaped like(32), act as(32), have(28), be called(28), consist of(27), be classified as(26), use(26), work as(24), emulate(23), be made of(22), behave like(21), appear(20), emerge from(17), concern(17), be about(15), serve as(15), require(14), feel like(14), be related to(14), appear like(13), be composed of(13), represent(12), mimic(11), become(11), remind of(11), be comprised of(10), be made up of(10), be considered(10), describe(10), be made from(10), belong to(10), swim like(10), relate to(10), be in(9), work like(9), cover like(9), seem(8), grow into(8), be described as(8), be based on(8), imitate(8), refer to(8), form(7), be from(7), revolve around(7), live like(7), be used as(6), exist as(6), utilize(6), hit(6), belong in(6), support(6), seem like(6), be for(6), be categorized as(6), behave as(6), pertain to(6), be born(6), be employed as(5), occur in(5), provide(5), smell like(5), focus on(5), define(5), be born of(5), recall(5), be named(5), incorporate(4), work with(4), live as(4), play(4), be created by(4), produce(4), make(4), be at(4), study(4), reach(4), supply(4), occur during(4), depend on(4), be of(4), accuse of(4), be found in(4), be on(4), derive from(4), be named after(4), descend from(4), encourage(3), encompass(3), push(3), like(3), affect(3), function as(3), be made by(3), be sold as(3), promote(3), smell of(3), hold(3), be defined as(3), be concerned with(3), rely on(3), portray(3), be thin like(3), be born from(3), be seen as(3), go in(3), simulate(3), copy(3), be birthed by(3), be shaped as(3), move like(3), be formed like(3), generate from(3), begin with(3), take after(3), be referred to as(3), center around(3), perform as(3), be supported by(2), substitute for(2), display(2), grow like(2), be formed by(2), move(2), accept(2), go into(2), inspire(2), specify(2), do(2), happen during(2), be identified as(2), comprise(2), teach(2), approach(2), exist in(2), mean(2), originate from(2), come in(2), accompany(2), house(2), help(2), ...

Web: be(3949), include(955), help(397), work with(276), provide(233), support(228), have(218), become(218), assist(210), look like(190), serve(167), resemble(156), represent(137), play(132), work as(127), bring(123), involve(112), make(95), serve as(95), range from(93), teach(91), educate(81), allow(75), form(72), enable(72), encourage(70), be at(67), recognise(62), reach(56), give(49), capture(48), study(48), be called(46), be considered(44), train(44), promote(42), consist of(40), use(39), require(38), connect(38), know(37), appear as(36), win(35), attempt(34), act as(33), lose(33), be in(32), be created by(32), portray(31), put(31), exceed(31), find(30), compete with(30), warm(28), specialize in(28), follow(28), empower(28), match(28), be like(27), occur in(26), want(26), take(26), be founded by(26), deal with(24), launch(24), link(24), reduce(24), live as(23), feature(22), offer(22), host(22), make up(21), contain(21), function as(21), be affiliated with(21), organize(21), recognize(21), be treated as(20), constitute(20), meet as(19), kill(19), join(19), lie at(19), guide(19), be run by(18), start as(18), surround(18), write(18), be brought up with(18), prevent(18), get(18), cut(18), tell(18), inspire(17), be comprised of(17), call(17), be during(17), lead(16), affect(16), accompany(16), comprise(16), hold(16), shoot(16), reflect(16), occur as(16), be formed by(15), be associated with(15), touch(15), benefit(15), rely on(15), treat(15), visit(15), coincide with(14), love(14), save(14), marry(14), advise(14), care about(14), engage(14), be started by(14), mimic(13), be classified as(13), mean that(13), advocate for(13), drive(13), inform(13), believe that(13), be used as(12), incorporate(12), focus on(12), meet with(12), increase(12), pose as(12), drive with(12), introduce(12), think(12), modify(12), sponsor(12), do(11), generate(11), qualify as(11), think that(11), be represented as(11), mobilize(11), run from(11), surpass(11), establish(11), name(11), dehumanize(11), murder(11), ...

0.67 CAUSE$_2$

Humans: be caused by(240), come from(117), result from(80), be created by(42), be due to(35), be made by(34), involve(22), be related to(21), emerge from(20), be induced by(14), occur because of(14), happen because of(12), be produced by(10), be associated with(10), be from(10), be derived from(9), occur in(8), be brought on by(8), occur during(8), stem from(7), be blamed on(7), be generated by(7), indicate(6), occur from(6), happen in(6), be left by(6), be instigated by(5), accompany(5), need(5), be because of(5), derive from(5), happen during(5), emanate from(5), be found in(5), be eaten by(5), contain(5), originate from(5), be made from(5), use(4), follow(4), look like(4), be made with(4), be produced from(4), start with(4), be triggered by(4), be started by(4), be affected by(3), resemble(3), cause(3), show(3), begin with(3), coincide with(3), arise from(3), be measured in(3), measure(3), control(3), be in(3), generate from(3), happen due to(3), concern(3), emit from(3), react to(3), bring on by(3), relate to(3), pertain to(3), occur due to(3), be about(3), develop from(2), come with(2), result in(2), precede(2), require(2), feed(2), be exacerbated by(2), be caused from(2), affect(2), create(2), produce(2), bespeak(2), be done with(2), form from(2), be stimulated by(2), predict(2), utilize(2), be linked to(2), keep out(2), develop during(2), be made of(2), originate with(2), happen with(2), occur at(2), be created from(2), be attributed to(2), concentrate(2), refer to(2), occur with(1), have(1), occur after(1), include(1), force(1), suggest(1), reveal(1), be(1), occur on(1), be aggravated by(1), announce(1), prepare for(1), happen from(1), come before(1), be inflicted by(1), be exerted by(1), come in(1), recognize(1), depend on(1), revolve around(1), keep inside(1), appear like(1), ostrasizes(1), come froom(1), begin by(1), worsen in(1), strike with(1), stain(1), be relative to(1), emanate(1), develop around(1), occur because(1), flow with(1), move with(1), appear during(1), be found with(1), be measured from(1), begin(1), be overfilled with(1), close with(1), develop by(1), be evidence of(1), come after(1), level to(1), be chewed by(1), come through(1), be sensitive to(1), protect(1), remain(1), inhibit(1), exude from(1), infect(1), prophecies(1), stain like(1), come on from(1), . . .

Web: be caused by(290), result from(217), cause(160), draw(109), be(97), be characterized by(81), have(76), be associated with(68), force(67), accompany(56), give(56), involve(54), result in(43), pull(41), arise from(39), prevent(34), drive(23), come with(22), make(19), suck(19), contaminate(18), bring(17), occur in(16), precede(16), come from(16), follow(16), be covered by(16), keep(15), push(15), include(14), attend(14), allow(14), deliver(14), blow(13), indicate(12), draw in(12), occur from(10), be exerted by(10), be suspended after(10), affect(8), move(8), stand for(8), reflect(8), supply(8), be removed with(8), reveal(7), cool(7), be fueled by(7), squeeze(7), occur with(6), create(6), depend on(6), move through(6), be removed by(6), occur without(6), be called(6), mention(6), be reversed by(6), exist in(6), mean(6), be governed by(6), be separated from(6), resemble(5), show(5), be blamed on(5), compress(5), indicate that(5), be inundated by(5), modify(5), herald(5), mean that(5), rise(5), accelerate(5), be diagnosed as(5), persist in(5), be characterised by(5), melt into(5), develop from(4), require(4), stem from(4), use(4), be aggravated by(4), be from(4), look like(4), depend upon(4), go with(4), set(4), be maintained in(4), pressurize(4), represent(4), propagate through(4), provide(4), attract(4), be observed with(4), condition(4), occur(4), be caused(4), chop(4), retain(4), be correlated with(4), displace(4), be related with(4), induce(4), be induced by(3), be exacerbated by(3), occur on(3), prepare for(3), ascend in(3), manifest with(3), get into(3), start as(3), drag(3), be hastened by(3), reverberate into(3), let(3), be on(3), leak(3), denser(3), wash away with(3), live in(3), be endured before(3), be marked by(3), stop(3), be accompanied by(3), remove(3), be caused be(3), surround(3), prompt(3), take(3), be recognized as(3), develop between(3), go away with(3), extend over(3), posit that(3), be washed out with(3), entrain(3), drag in(3), occur after(2), be affected by(2), feed(2), coincide with(2), be made by(2), produce(2), recognize(2), divert(2), exist between(2), score(2), push inside(2), consist in(2), be supplied through(2), . . .

0.59 NOMINALIZATION:AGENT

Humans: give(30), avoid(22), sort(18), donate(16), cut(16), organize(13), design(11), edit(11), work with(10), plan(10), supply(10), run from(9), slice(9), work for(8), evade(7), take(7), divide(6), provide(6), separate(5), hide from(5), handle(5), escape(5), count(4), elude(4), be concerned with(4), run away from(4), work on(4), arrange(4), lay out(3), collect(3), build(3), create(3), envision(3), make(3), administer(3), complete(3), record(3), trim(3), classify(3), compile(3), process(3), develop(2), come from(2), study(2), share(2), engineer(2), poll for(2), measure(2), ignore(2), improve(2), enhance(2), tally(2), finish(2), involve(2), think about(2), run through(2), participate in(2), go through(2), duck(2), splice(2), object to(2), snip(2), conduct(2), categorize(2), oversee(1), help(1), work in(1), choose(1), offer(1), dodge(1), deal with(1), analyze(1), carve(1), sever(1), contribute(1), chop up(1), manage(1), be employed by(1), defy(1), volunteer(1), write(1), answer(1), tend to(1), file(1), razor(1), depart with(1), regulate(1), divide up(1), stay away from(1), remove excess(1), collect data for(1), dislike(1), hand over(1), reorder(1), enumerate(1), dodge away from(1), shirk from(1), slice through(1), be used on(1), flee(1), expedite(1), prepare(1), proofread(1), accumulate(1), disapprove of(1), veer from(1), spurn(1), modify(1), operate on(1), divide apart(1), be made for(1), be in(1), control(1), eliminate(1), give out(1), help replenish supply of(1), be used with(1), map out(1), catagorizes(1), escape from(1), defer(1), collate(1), cut apart(1), be designed for(1), get out of(1), monitor(1), oppose(1), chop(1), blueprint(1), leak(1), visualize(1), partition(1), design of(1), elude from(1), dispense(1), manage for(1), discuss(1), endorse(1), plot(1), plan for(1), be afraid of(1), go into(1), resize(1), distribute(1), alter(1), shorten(1), pertain to(1), specialize in(1), release(1), brainstorm(1), ...

Web: give(654), donate(395), receive(74), sell(41), provide(40), work with(27), sort(25), work for(25), serve on(22), cut(20), supply(17), be(15), work in(14), design(13), treat(12), build(11), match(11), contribute(10), offer(9), have(9), work at(9), mean(9), see(9), donate whole(8), visit(8), dodge(7), deal with(7), include(7), understand(7), believe that(7), evade(6), lay out(6), create(6), manage(6), write(6), look at(6), view(6), develop(5), divide(5), choose(5), plan(5), study(5), restructure(5), conceive of(5), consider(5), give whole(5), transform(5), craft(5), giv(5), be employed by(4), meet(4), coordinate with(4), saw(4), tour(4), love(4), represent(4), promote(4), exclude(4), be with(4), sweep(4), work between(4), qualify as(4), serve(4), be transmitted by(4), want(4), coordinate(4), separate(3), take(3), chop up(3), envision(3), advise(3), value(3), head(3), remake(3), fight for(3), foster(3), conform(3), deplore(3), favor(3), think that(3), candonate(3), believe(3), destroy(3), be vaccinated for(3), be given(3), be of(3), be told(3), consult with(3), be hired by(3), come from(2), carve(2), sever(2), share(2), defy(2), be recruited by(2), occur in(2), care for(2), ruin(2), incorporate(2), identify(2), protest(2), perforate(2), be exposed from(2), lobby(2), be housed within(2), be located within(2), transform of(2), argue that(2), realize that(2), assist(2), designate(2), be infected with(2), take over(2), be involved in(2), staff(2), pass(2), pick(2), be spiked with(2), mean that(2), know(2), need(2), tell(2), urge(2), rewrite(2), be employed in(2), enter(2), selflessly(2), serve as(2), allow(2), criticize(2), oversee(1), help(1), avoid(1), analyze(1), collect(1), volunteer(1), join(1), book(1), draw(1), estimate that(1), act in(1), indicate that(1), result in(1), forget(1), prosecute(1), spare(1), disclose(1), operate in(1), experiment with(1), write on(1), ...

0.55 USE

Humans: use(140), utilize(46), require(37), involve(29), come from(28), need(22), cook with(18), consist of(17), be done by(17), be made of(17), be made by(15), create(15), contain(14), run on(13), be operated by(12), cause(12), employ(12), work with(11), be used by(11), rely on(11), be made from(10), work by(9), emit(9), produce(9), be powered by(9), be played with(8), burn(8), be worn on(8), be composed of(8), originate from(7), be caused by(7), include(6), move(6), be produced by(6), be activated by(6), make(6), be performed by(6), depend on(6), emerge from(6), be played by(6), be comprised of(6), clean by(6), have(5), apply(5), be(5), resemble(5), work on(5), lead to(5), come by(5), result in(5), be based on(5), incorporate(4), be controlled by(4), operate by(4), be made for(4), induce(4), look like(4), result from(4), emanate from(4), consume(3), fuse(3), generate(3), be by(3), be applied by(3), be generated by(3), release(3), attach to(3), be developed by(3), be conducted by(3), be completed by(3), be related to(3), give off(3), cook use(3), be created by(3), be given by(3), be made up of(3), rely upon(3), depend upon(3), cook under(3), cook by(3), be cast using(3), necessitate(3), work in(2), mimic(2), be fueled by(2), entail(2), be broadcast by(2), be used with(2), be provided by(2), be heated by(2), modify(2), spin(2), go on(2), give(2), be like(2), begin with(2), ignite(2), be worked by(2), function with(2), count(2), encourage(2), harness(2), be brought about by(2), be used on(2), be made with(2), happen through(2), be cast by(2), go into(2), stop by(2), flow through(2), be held by(2), heat by(2), be placed on(2), be made using(2), happen over(2), rotate in(2), utilizes(2), heat with(2), be due to(2), build up(2), be created from(2), be carried by(2), turn by(2), be emitted by(2), bring about(2), turn use(2), occur in(2), be induced by(2), generate from(2), concern(2), be counted by(2), be run by(2), pass through(2), happen because of(2), be handled by(2), happen by(2), run by(2), believe in(2), occur because of(2), operate with(2), relate to(2), transform(2), occur through(2), be transmitted by(2), suck into(2), feature(2), spray(1), bring(1), carry(1), be taken by(1), ...

Web: be(160), use(150), be operated by(112), provide(47), include(46), receive(44), raise(44), run on(38), burn(33), lift(32), pump(30), have(28), bring(28), make(27), be turned by(26), produce(24), draw(24), involve(22), spill(20), allow(18), generate(17), be broadcast by(14), be applied by(14), require(13), be powered by(13), drive(12), invent(12), give(11), deliver(11), be heated by(10), be transmitted over(10), prevent(10), cause(9), represent(9), operate at(9), convert(9), churn(9), maintain(8), be like(8), own(8), seem like(8), work with(7), take(7), be used with(7), discharge(7), need(7), rely on(7), run over(7), be wrought by(7), carry(6), move(6), leak(6), be controlled by(6), consume(6), operate on(6), contain(6), look like(6), absorb(6), be under(6), scoop(6), reduce(6), develop(6), supply(6), be run with(6), appear on(6), work in(5), be fueled by(5), be produced by(5), do(5), utilize(5), emit(5), occur over(5), resemble(5), employ(5), be used as(5), mean that(5), be called(5), be released by(5), dominate(5), tend toward(5), be rejected by(5), pressurize(5), establish(5), smell of(5), pick up than(5), be set by(5), sell(5), squeeze(5), purify(5), empty(5), be driven by(5), remove(5), enable(5), be accomplished by(4), simulate(4), comprise(4), be activated by(4), fuse(4), be filled with(4), be performed by(4), be by(4), draw up(4), be corrected by(4), control(4), lift up(4), convey(4), design(4), combine(4), rest in(4), run(4), mean(4), accompany(4), turn off(4), admit(4), detonate(4), be behind(4), require that(4), work like(4), turn(4), spray(3), mimic(3), apply(3), be designed for(3), turn with(3), be altered by(3), repair(3), be generated by(3), be worked by(3), depend on(3), throw(3), be mounted on(3), advocate(3), ignore(3), pass by(3), operate(3), be obtained by(3), cook(3), slosh(3), be influenced by(3), secure(3), splash(3), be known by(3), notify(3), be of(3), be imported in(3), live(3), permit(3), spit(3), be propelled in(3), understand(3), ...

0.54 HAVE$_1$

Humans: contain(157), have(125), be made from(52), include(50), be made of(47), be made with(33), use(33), be composed of(23), consist of(22), bear(20), taste like(19), produce(19), grow(15), own(14), come from(14), display(12), house(10), feature(10), utilize(10), provide(9), show(9), love(9), supply(9), be made up of(9), taste of(8), be in(8), possess(8), incorporate(7), be filled with(7), be(7), support(7), hold(7), be inhabited by(6), care for(6), be made using(6), make(5), resemble(5), give(5), surround(5), look like(5), be flavored with(5), harbor(5), be comprised of(5), display in(5), require(4), enjoy(4), like(4), have within(4), employ(4), host(4), be created from(4), be trimmed with(3), feed(3), involve(3), be decorated with(3), sustain(3), depend on(3), rely on(3), be baked with(3), smell of(3), showcases(3), be of(3), be prepared with(3), be made out of(3), nourish(3), project in(3), nurture(3), broadcast in(3), keep(2), protect(2), be like(2), be near(2), shelter(2), be based on(2), be mixed with(2), develop(2), be garnished with(2), form(2), concentrate(2), look after(2), be about(2), accommodate(2), be derived from(2), be cooked with(2), be populated by(2), show in(2), project(2), transmit in(2), be found in(2), be populated with(2), be constructed from(2), be capable of(2), perform(2), be trimmed in(2), be used as(1), live with(1), adopt(1), foster(1), be infested with(1), be laden with(1), call for(1), watch(1), present(1), destroy(1), add(1), need(1), be grown for(1), be known for(1), extend(1), benefit from(1), allow(1), be saturated with(1), bare(1), permit(1), start with(1), be centered around(1), center around(1), offer(1), collect(1), boast of(1), grow from(1), retain(1), create(1), deliver(1), evoke(1), play in(1), be illustrated with(1), welcome(1), belong to(1), encourage(1), caintains(1), depend upon(1), be blended with(1), be created out of(1), be constructed of(1), remind of(1), be created with(1), comtains(1), issue(1), blossom with(1), be smooth like(1), add on(1), sink under(1), encompass(1), bloom(1), emerge from(1), emanate(1), result in(1), be focused on(1), cook up with(1), ...

Web: bear(2114), have(1885), produce(1230), contain(441), include(435), give(428), be(383), bore(269), provide(222), yield(203), show(201), bring(194), bringeth(140), paint(119), use(89), lose(87), adopt(85), grow(76), present(68), feature(68), carry(67), offer(53), drop(53), be loaded with(47), love(46), want(46), be filled with(43), own(38), keep(38), set(37), do(36), combine(32), take(31), beareth(27), be made with(26), consist of(26), be in(25), make(23), hold(22), get(21), need(20), lack(19), be illustrated with(18), fill(17), supply(15), draw(15), describe(15), give up(15), taste like(14), bare(14), look for(14), have in(14), become(14), allow(13), house(13), know(13), put(13), form(12), protect(12), travel with(12), introduce(12), attend(11), furnish(11), put forth(11), be called(11), cause(11), cast(10), be than(10), look at(10), be made of(9), rely on(9), range from(9), afford(9), be used as(8), be laden with(8), like(8), collect(8), be like(8), create(8), treat(8), leave(8), take in(8), be packed with(8), consider(8), preserve(8), be located within(8), dissect(8), shed(8), miss(8), call for(7), care for(7), host(7), be grown for(7), be composed of(7), support(7), retain(7), be made from(7), abandon(7), throw(7), bring forth(7), be used in(7), puree(7), be harvested for(7), enjoy(6), be based on(6), be mixed with(6), grow from(6), look after(6), hurt(6), sprout(6), teach with(6), rotten(6), dry(6), register(6), maintain(6), be known as(6), find(6), foster(5), display(5), feed(5), add(5), start with(5), look like(5), move(5), match(5), care about(5), inspire(5), be used for(5), be ornamented with(5), live(5), interpret(5), raise(5), avoid(5), be on(5), dissolve(5), appear in(5), be cultivated for(5), hang(5), colour(5), be enriched with(5), go(5), be stripped of(5), invent(5), incorporate(4), welcome(4), resemble(4), develop(4), surround(4), center around(4), concentrate(4), ...

0.53 FROM

Humans: come from(281), be made from(139), be made of(64), contain(47), be produced from(35), be derived from(31), be from(31), be made in(26), originate from(25), taste like(23), work on(23), live on(22), be composed of(22), live in(21), be found in(19), consist of(19), emerge from(18), derive from(18), be created from(18), use(13), be conceived in(13), be made at(12), be used in(11), be(11), be sold in(11), be born in(10), be comprised of(10), go to(10), be distilled from(9), smell like(9), be produced by(9), originate in(9), be made with(8), resemble(8), visit(8), be bought in(8), be pressed from(8), be created in(8), be made out of(8), be created by(8), be processed from(8), include(7), be rendered from(7), travel to(7), taste of(6), be used at(6), be raised in(6), be produced in(6), be extracted from(6), result from(6), be bred in(6), be taken from(5), require(5), be manufactured from(5), be purchased at(5), be associated with(5), look like(5), begin in(5), be in(4), blow from(4), be found at(4), be squeezed from(4), hail from(4), belong in(4), come out of(4), be purchased in(4), be found on(4), be formulated from(4), be crushed from(4), be made using(4), be started in(4), be generated by(4), be brewed from(4), be worn in(4), be fertilized in(4), be sold by(3), reside on(3), work in(3), be obtained from(3), be generated from(3), enjoy(3), be born on(3), be bought from(3), tour(3), be fermented from(3), generate from(3), involve(3), be caused by(3), be near(3), start in(3), be located in(3), emanate from(3), originate at(3), be purchased from(3), make(2), represent(2), need(2), smell of(2), come off(2), be produced with(2), reside in(2), be of(2), flow from(2), be raised on(2), grow up on(2), blow off(2), stay in(2), belong to(2), be created out of(2), remind of(2), explore(2), toil on(2), be grown in(2), be contained in(2), recall(2), be pressed out of(2), be made for(2), be lived in(2), be shed from(2), be adopted from(2), be formed from(2), be fashioned from(2), feel like(2), emit from(2), stem from(2), be churned in(2), stick to(2), be formed in(2), be based on(2), make of(2), be produced using(2), be related to(2), be bought at(2), labor on(2), reside at(2), come off of(2), be housed in(2), inhabit(2), be crafted from(2), happen at(2), cover(2), be developed at(2), descend from(2), be made up of(2), sound like(2), substitute for(1), live near(1), come with(1), buy(1), travel from(1), ...

Web: come from(204), live on(175), be(146), enter(134), be in(113), pee in(110), leave(107), grow up on(103), blow from(72), work on(72), have(57), contain(53), visit(49), be made from(42), include(38), come into(34), be from(32), live in(29), be sold in(29), resemble(26), arrive in(24), be extracted from(23), make(22), taste like(21), reside in(21), be sold as(21), arrive at(21), stay in(18), arrive from(17), be prepared at(16), be obtained from(16), come off(15), be produced from(15), cause(15), remain on(13), blow towards(13), ruffle(13), be substituted for(13), use(12), be raised on(12), be kept in(12), be derived from(12), be produced in(12), be raised in(11), be tried at(11), sell in(11), raise(11), go from(11), be pressed from(10), be made at(10), work(10), descend on(10), produce(10), remain in(10), be found in(9), tour(9), taste of(9), love(9), frilled(9), offer(9), hit(9), occur in(9), be reared on(9), come over(8), travel in(8), be used in(8), be produced with(8), fall from(8), ebb over(8), rival(8), sell(8), rise with(8), compete with(8), be born on(7), hail from(7), rise from(7), know(7), tame(7), be refined(7), work at(7), turn(7), be termed(7), require(6), be manufactured from(6), represent(6), be sold by(6), buy(6), flow from(6), be done at(6), look like(6), gambol over(6), show(6), take(6), replace(6), view(6), be sold at(6), be blended with(6), be at(6), divide(6), be pressed(5), be associated with(5), smell of(5), be made of(5), fit in(5), predominate over(5), go into(5), be marketed as(5), fly from(5), run(5), write about(5), be labeled(5), take over(5), give(5), fuel(5), inherit(5), hold(5), evoke(5), press(5), reside on(4), be distilled from(4), be purchased at(4), be squeezed from(4), like(4), run from(4), be administered at(4), be used at(4), be on(4), smell like(4), be generated from(4), be born in(4), be rendered from(4), domesticate(4), build(4), be wafted across(4), fly over(4), stumble upon(4), travel through(4), be packed with(4), travel between(4), consider(4), result in(4), compare with(4), be than(4), revitalize(4), stay on(4), pass through(4), be mixed with(4), be bought by(4), saw(4), be trained on(4), ...

0.47 FOR

Humans: contain(64), be for(33), be used for(31), treat(26), hold(26), dispense(23), produce(20), attract(19), sell(19), cure(19), make(17), have(15), feed(15), manufacture(15), help(15), assemble(15), include(14), nourish(14), heal(14), try(13), be used in(13), deal with(13), display(12), build(12), be used on(11), involve(11), work with(10), work in(9), relieve(9), medicate(9), study(9), store(8), protect(8), specialize in(8), be made for(8), be meant for(8), vend(7), aid(7), feature(7), protect from(7), pay for(7), draw(6), conduct(6), provide(6), know about(6), care for(6), distribute(6), handle(5), buy(5), focus on(5), reduce(5), come from(5), pump(5), supply(5), serve(5), sustain(5), judge(5), channel(5), concern(5), be given to(5), relate to(5), be applied to(5), facilitate(4), affect(4), go in(4), be intended for(4), tend to(4), improve(4), clear(4), soothe(4), be placed in(4), cover(4), be concerned with(4), work on(4), fix(4), prosecute(3), direct(3), enhance(3), go into(3), show(3), do(3), be used by(3), offer(3), process(3), house(3), be filled with(3), give(3), help with(3), give out(3), create(3), adjudicate(3), be worn by(3), showcases(3), design(3), be needed for(3), organize(3), deflect(3), be fed to(3), ease(3), show off(3), examine(3), divert(2), encompass(2), spew(2), purchase(2), know(2), sentence(2), use(2), be utilized in(2), extract(2), understand(2), be found in(2), gush(2), be required for(2), keep(2), be used during(2), resemble(2), benefit(2), turn out(2), stop(2), be made of(2), advertise(2), release(2), knock out(2), decrease(2), be inserted into(2), be put in(2), be sprayed on(2), demonstrate(2), be designed for(2), absorb(2), provide for(2), manage(2), do away with(2), be employed in(2), moisturize(2), trap(2), be comprised of(2), exhibit(2), put together(2), learn about(2), dissipate(2), pertain to(2), run from(2), construct(2), investigate(2), repel(2), alleviate(2), go on(2), serve for(2), ...

Web: produce(709), have(375), contain(349), include(260), be(167), come from(155), find(85), be derived from(74), give(63), test(58), be used for(49), pump(44), interrupt(43), need(43), cause(41), kill(38), show(36), hold(36), feed(34), flow(34), allow(34), be filled with(33), be drilled for(32), feature(31), keep(30), make(30), state that(29), sell(27), operate(25), yield(22), place(22), target(20), be used in(20), encounter(20), display(19), strike(19), commit(18), attract(17), consist of(16), handle(16), dispense(16), specialize in(16), discover(16), cure(15), deal with(15), hit(15), be utilized in(14), paint(14), build(13), nourish(13), adjudicate(13), be struck by(13), draw(12), deter(12), sit as(12), identify(12), use(11), push(11), don(11), work in(10), be used by(10), require(10), be designated as(10), prevent(9), do(9), offer(9), develop(9), release(9), ride(9), be melted by(9), convey(9), own(9), drive(9), harm(9), present(9), be known as(9), confirm(9), be completed as(9), take away(9), be produced by(8), play(8), resemble(8), provide(8), supply(8), be made from(8), turn away(8), take(8), recover(8), travel by(8), receive(8), be from(8), see(8), be administered by(8), treat(7), manufacture(7), allow for(7), look like(7), bring(7), be ripened on(7), carry(7), be produced in(7), be in(6), work for(6), try(6), put(6), stop(6), enable(6), serve(6), tap(6), leak(6), stay within(6), run down(6), combine(6), capture(6), drill for(6), be provided by(6), modify(6), be obtained from(6), prove(6), work with(5), purchase(5), be associated with(5), relieve(5), house(5), assemble(5), sustain(5), dislike(5), pump out(5), be suited for(5), apply throughout(5), comprise(5), be made by(5), reverse(5), hear(5), go with(5), produce as(5), reveal(5), grip(5), adjudge(5), turn(5), transmit(5), be used on(4), drain(4), sentence(4), reduce(4), be for(4), serve in(4), come with(4), insure(4), force(4), ...

0.42 NOMINALIZATION:PRODUCT

Humans: be made by(47), involve(30), contain(25), come from(25), include(22), be(18), ask for(17), be for(17), concern(15), be made for(13), consist of(13), request(13), describe(12), be about(11), refer to(10), pay for(8), affect(8), shorten(8), be given by(8), decorate(7), be on(7), provide(7), be of(7), be made of(7), be made up of(7), require(6), supply(6), be comprised of(6), cut(6), purchase(5), appear on(5), define(5), deal with(5), fund(5), pertain to(5), be made from(5), be written by(4), cover(4), support(4), help with(4), be related to(4), be created by(4), bring in(4), be made about(4), explain(4), subsidize(4), have(3), reduce(3), pay(3), comprise(3), lower(3), emerge from(3), be decided by(3), tell about(3), be rendered by(3), happen to(3), depict(3), illuminate(3), beautify(3), neaten(3), be composed of(3), be done by(3), list(3), relate to(3), grace(2), distribute(2), be derived from(2), use(2), be used for(2), adorn(2), create(2), produce(2), enhance(2), be done to(2), control(2), clean up(2), demand(2), discuss(2), provide for(2), assist with(2), stem from(2), be developed by(2), focus on(2), be found on(2), cut back(2), be performed by(2), catalog(2), detail(2), enumerate(2), be located in(2), be given to(2), be by(2), be found in(2), be placed on(2), shape(2), characterize(2), trim(2), be from(2), originate from(2), requisition(2), be created for(2), result in(1), help(1), make(1), buy(1), be reserved for(1), maintain(1), be in(1), go on(1), represent(1), benefit(1), offset(1), want(1), be based on(1), utilize(1), influence(1), limit(1), form(1), be made in(1), be shared between(1), hold(1), serve(1), look like(1), revolve around(1), style(1), comment on(1), be preformed on(1), go around(1), be shared by(1), redistribute(1), about(1), come out of(1), be comprised by(1), emanate(1), decrease(1), wipe out(1), judge(1), be offered by(1), become against(1), improve(1), tidy up(1), be engendered by(1), modify(1), be formed by(1), be created as(1), shear(1), be formulated by(1), be made to(1), jacket(1), call for(1), ...

Web: include(99), allow(93), affect(65), be(58), appear on(42), make(36), be made by(36), grace(33), have(24), provide(19), be used for(17), consist of(16), become(16), be on(15), feature(13), help(11), require(10), involve(10), do(10), limit(9), be captured on(9), design(9), go against(9), utilize(8), reduce(8), find(8), maintain(7), keep(7), increase(7), be featured on(7), feature on(7), double as(7), be derived from(6), use(6), be for(6), push(6), mean that(6), be criticized by(6), adorn(5), comprise(5), create(5), be used as(5), be supported by(5), incorporate(5), embrace(5), rest with(5), constitute(5), restrict(5), flow around(5), distinguish(5), deploy(5), result in(4), pay for(4), be made for(4), lower(4), be about(4), donate(4), eat(4), handle(4), be used on(4), secure(4), find that(4), increase for(4), be sent between(4), minimize(4), enable(4), differ from(4), range from(4), lead(4), impact(4), be dispersed among(4), brace(4), be swiped from(4), engage(4), decorate(3), be of(3), influence(3), be shared between(3), hold(3), define(3), encourage(3), bring(3), motivate(3), coordinate with(3), dismay(3), be dominated by(3), be incorporated into(3), complement(3), be purchased through(3), present(3), let(3), exclude(3), be located under(3), indulge in(3), be approved by(3), leave(3), handwritten(3), embellish(3), publish(3), require that(3), replace(3), open(3), mean(3), preclude(3), place(3), be faced by(3), ask for(2), be written by(2), represent(2), benefit(2), offset(2), want(2), be based on(2), form(2), contain(2), come from(2), serve(2), produce(2), package(2), jeopardize(2), render(2), ensure at(2), concur with(2), relate(2), approximate(2), stun(2), be stamped on(2), intoxicate(2), flop(2), be mounted under(2), appear in(2), borrow(2), be fed through(2), traverse(2), be ironed out at(2), permit(2), send(2), span(2), remove(2), divide(2), appear upon(2), bias(2), reflect(2), offer(2), run against(2), be depicted on(2), teach(2), hinge(2), double after(2), extend across(2), ...

0.34 IN

Humans: live in(68), occur in(61), be in(49), occur during(48), come from(46), be found in(38), happen in(38), involve(32), reside in(31), occur at(30), happen at(26), happen during(25), be located in(23), occur on(20), be eaten at(19), inhabit(18), affect(18), be played in(17), be said in(12), be consumed at(11), happen on(11), be(10), exist in(10), be on(10), start in(10), be believed by(10), begin in(9), be experienced by(9), originate in(9), emerge from(8), dwell in(8), belong to(8), relate to(8), be built in(7), include(7), be caused by(7), be made in(7), happen to(7), be related to(7), be during(6), be of(6), begin at(6), come in(6), plague(6), be done in(6), support(6), be held by(6), pertain to(6), be from(5), concern(5), occupy(5), divide(5), make up(5), define(5), be made by(5), belong in(5), be written by(5), result from(4), require(4), start(4), be recited in(4), be played during(4), stem from(4), survive in(4), occur within(4), occur(4), be situated in(4), fly at(4), be taken at(4), form in(4), comprise(4), be provided by(4), come at(4), be found on(4), be situated on(4), refer to(4), go into(3), be followed by(3), be within(3), be created by(3), be with(3), come on(3), revolve around(3), destroy(3), be scheduled for(3), be near(3), benefit(3), appear in(3), arise in(3), surround(3), deal with(3), characterize(3), effect(3), serve(3), thrive in(3), trouble(3), be sponsored by(3), enhance(3), go at(3), be formed in(3), be due to(3), originate from(3), found in(3), be taken in(3), hide in(3), feed in(3), have(2), control(2), operate(2), consist of(2), happen(2), represent(2), feed(2), afflict(2), befall(2), bother(2), run in(2), use(2), be solved by(2), utilize(2), exist on(2), be said(2), populate(2), exist within(2), stay in(2), accompany(2), be considered(2), be about(2), leave at(2), disrupt(2), result in(2), remain in(2), be suffered by(2), cause(2), breed in(2), change(2), challenge(2), begin during(2), contain(2), cover(2), create(2), run through(2), start at(2), be supported by(2), be eaten around(2), happen near(2), reside on(2), be offered by(2), shatter(2), celebrate(2), . . .

Web: include(603), affect(319), live in(302), be(141), plague(128), bring(108), arise in(90), confront(79), lie beneath(69), be in(62), occur in(61), cause(61), run in(50), come from(49), have(47), involve(47), simmer beneath(44), leave(40), make(33), threaten(33), exist in(31), lurk beneath(30), face(29), extend into(29), leave at(28), move from(28), lead(28), arise within(28), run(28), beset(27), follow(27), keep(26), destroy(24), do(24), force(23), impact(23), tear(23), put(23), lay beneath(23), begin in(22), simmer under(22), be faced by(20), be under(20), represent(19), hold(19), devastate(18), be adopted by(18), be below(18), work in(17), divide(17), flee(17), simmer below(17), be from(16), occur within(16), know(16), arise from(16), be held(16), clean(16), break(15), bubble beneath(15), result in(14), concern(14), work(14), work for(14), exist for(14), be during(13), populate(13), characterize(13), push(13), mean(13), exist beneath(13), arise for(13), be read(13), come into(13), require(12), start(12), be solved by(12), boil beneath(12), haunt(12), touch(12), exist below(12), disrupt(11), be at(11), make up(11), love(11), inhabit(11), be considered(11), arrive at(11), dwell in(11), serve(11), live outside(11), build(11), drive(11), seethe beneath(11), be set by(11), ravage(11), grow up in(11), allow(11), give(11), be offered(11), be recited(10), reflect(10), exist within(10), run through(10), create(10), burden(10), inhibit(10), prevent(10), end(10), take(10), ensnare(10), exist at(10), provide(10), be beneath(10), hurt(9), be called(9), be said(9), afflict(9), happen in(9), challenge(9), be generated by(9), mean that(9), run into(9), stretch into(9), mark(9), visit(9), be located in(8), start in(8), bubble below(8), open in(8), remain under(8), lie below(8), prevail in(8), dog(8), volk for(8), suggest that(8), use(7), effect(7), constitute(7), be played in(7), occur at(7), arrive(7), lie under(7), be encountered in(7), know that(7), reduce(7), remain beneath(7), exist underneath(7), strain(7), care about(7), . . .

0.30 NOMINALIZATION:PATIENT

Humans: be made by(43), come from(30), work for(28), be employed by(18), be created by(17), work at(15), work in(11), be designed by(11), be produced by(10), be from(8), involve(7), be trained by(7), be developed by(5), be hired by(5), be invented by(5), be paid by(5), be in(4), live in(4), be made in(4), learn from(3), be used by(3), be trained in(3), originate from(3), be thought of by(3), labor for(3), be presented by(2), run(2), be(2), belong to(2), be trained for(2), be conceived by(2), be rejected in(2), go to(2), contribute to(2), train in(2), be imagined by(2), be discarded by(2), challenge(1), help(1), attend(1), graduate from(1), staff(1), be done by(1), be enrolled in(1), serve(1), teach at(1), be retained by(1), study at(1), enter(1), serve at(1), be performed by(1), commence from(1), occur in(1), dissatisfy(1), be constructed by(1), make by(1), learn in(1), be disposed off by(1), earn from(1), be generated in(1), be built by(1), come out of(1), transpire in(1), be thought up by(1), misrepresent(1), emerge from(1), be apprenticed at(1), be determined in(1), be originated in(1), be credited to(1), dissapoint(1), go down in(1), flow from(1), be displayed(1), be devised by(1), happen in(1), be out of(1), be denied by(1), be by(1), emanate from(1), be formed by(1), be decided by(1), be found in(1), be judged by(1), be envisioned by(1), be eliminated by(1), be sketched by(1), inspire by(1), be made for(1), originate with(1), lose function in(1), be refused from(1), advise at(1), be shunned from(1), apprentice in(1), be generated by(1), happen(1), overpay(1), be inspired by(1), be sewn by(1), govern(1), be dreamt by(1), be located at(1), be matriculated at(1), be trained about(1), provide work for(1), be rejected by(1), study(1), originate in(1), originate at(1), be fired from(1), be associated with(1), produce by(1), be engineered by(1), be based in(1), be discoved by(1), be disposed of by(1), be derived from(1), be taught by(1), be dismissed by(1), be destroyed in(1), embarrass(1), be of(1), be disallowed by(1), focus on(1), pass through(1), toil at(1), appear in(1), plague(1), be payed by(1), be implemented by(1), be maunfactured by(1), help out at(1), train with(1), supply work for(1), be educated by(1).

Web: leave(98), be employed by(72), graduate from(63), attend(50), work for(32), work at(29), serve(29), have(23), be enrolled in(22), retire from(22), join(19), allow(18), violate(11), resign from(11), work in(10), be(10), be employed at(10), separate from(10), be made by(9), complete(9), inform(8), wish(8), notify(8), support(7), be employed with(7), disregard(6), perceive(6), enable(6), access(6), represent(6), be with(6), participate in(6), terminate from(6), travel for(6), come from(5), be produced by(5), affect(5), be in(4), be developed by(4), be performed by(4), enroll in(4), be required by(4), sue(4), be at(4), live outside(4), be selected with(4), maximize(4), be terminated by(4), separate under(4), be found by(4), be covered under(4), be created by(3), staff(3), enter(3), serve at(3), begin(3), take(3), finish(3), require(3), be insured by(3), be relieved from(3), be designated by(3), be funded by(3), start(3), graduate(3), possess(3), be discussed by(3), do(3), work within(3), be reemployed by(3), authorize(3), reduce(3), be done by(2), be from(2), be retained by(2), be hired by(2), be utilized by(2), relocate within(2), lack(2), meet(2), supervise(2), get(2), reimburse(2), work with(2), be authorized by(2), agree that(2), move(2), assist(2), enrol at(2), be involved in(2), focus(2), aid(2), pursue(2), land(2), drive(2), aide(2), accredit(2), be laid off by(2), put through(2), take up(2), be near(2), design(2), show(2), participate(2), transitioning from(2), use(2), be separated from(2), attain(2), be accredited at(2), stimulate(2), challenge(1), help(1), be presented by(1), run(1), involve(1), live in(1), teach at(1), study at(1), be split across(1), oversee(1), be nominated by(1), include(1), transfer from(1), operate(1), steal from(1), capture(1), present(1), pass(1), love(1), be on(1), seek(1), care about(1), be educated at(1), serve with(1), save(1), fill(1), remain with(1), save for(1), be employed within(1), become at(1), be employed in(1), enrol in(1), work(1), be crated by(1), please(1), be reimbursed by(1), benefit from(1), exemplify(1), be concerned that(1), leave for(1), . . .

0.30 MAKE$_2$

Humans: be made of(256), contain(115), be made from(97), be composed of(93), consist of(82), be comprised of(52), be(42), be made up of(38), include(30), have(21), look like(21), be manufactured from(21), come from(20), be formed from(19), be created from(18), taste like(17), resemble(16), use(16), involve(12), be made out of(12), house(11), be formed of(10), utilize(9), be constructed from(9), be cast from(7), be in(7), comprise(7), be constructed of(6), be created with(6), be formed by(6), be fashioned from(6), be printed on(6), emerge from(6), appear(6), be built from(6), be made with(5), be filled with(5), be of(5), belong to(5), be produced from(5), be inhabited by(5), be derived from(5), feel like(5), be lived in by(5), be fabricated from(5), incorporate(4), be on(4), represent(4), be run by(4), be built of(4), employ(4), hold(4), work on(4), be created out of(4), be constructed out of(4), be constituted by(4), be made using(4), be molded from(4), relate to(4), encompass(3), be used on(3), form from(3), link(3), require(3), provide(3), empower(3), run on(3), cut(3), support(3), be from(3), descend from(3), feature(3), be owned by(3), be formulated from(3), be created by(3), originate from(3), protect(2), be cast of(2), be like(2), be found in(2), be linked by(2), shape(2), connect(2), be cast in(2), run through(2), be eaten like(2), carry(2), taste of(2), be shaped like(2), pile up(2), be formulated of(2), smell of(2), be watched on(2), emulate(2), begin with(2), possess(2), be formed in(2), create from(2), be sculpted in(2), fold like(2), be occupied by(2), be formed with(2), be carved from(2), store(2), be for(2), melt into(2), be connected by(1), bring(1), be viewed on(1), start as(1), be joined by(1), be supplied by(1), work with(1), be replaced by(1), judge(1), come through(1), drip(1), be represented by(1), compose of(1), be protected by(1), be covered in(1), result from(1), turn into(1), need(1), be defended by(1), provide for(1), train(1), be covered with(1), sell(1), smell like(1), comprise of(1), benefit(1), own(1), be found on(1), be carried on(1), be composed from(1), be staffed by(1), be shown on(1), form(1), enable(1), be led by(1), affect(1), seem like(1), create(1), select(1), deliver(1), be made by(1), be struck in(1), be concerned with(1), serve(1), . . .

Web: include(1037), be(552), consist of(261), use(161), have(146), allow(119), contain(88), be made of(83), be of(79), be comprised of(76), be composed of(72), meet with(67), select(64), provide(63), advise(62), assist(62), guide(60), bring(57), affect(56), work with(54), evaluate(54), represent(53), need(53), be cast in(52), use in(52), require(51), review(46), involve(36), supervise(35), comprise(34), serve(34), examine(31), encourage(29), cause(28), make(27), oversee(26), give(26), inform(25), be made from(24), approve(23), wear(23), place(21), help(21), connect(20), recommend(20), look like(19), run(19), seek(19), become(19), run over(18), look through(18), employ(17), interview(17), nominate(17), be hidden behind(17), taste like(16), hear(16), remove(16), compete with(16), match(16), recognize(16), admit(16), be chosen by(16), be corrected with(16), exist on(15), consider(15), be in(14), combine(14), look for(14), choose(14), support(13), cut(13), fall into(13), be used in(13), break(13), be chaired by(13), compete in(13), educate(13), benefit(12), promote(12), move(12), rely on(12), operate(11), be fitted with(11), lead(11), be spent on(11), address(11), accept(11), lay(11), consult with(11), be broadcast via(11), be based on(11), incorporate(10), be made with(10), exclude(10), replace(10), welcome(10), put(10), assign(10), determine whether(10), be corrected by(10), be replaced with(10), bring together(10), be covered with(9), prepare(9), serve as(9), identify(9), act as(9), operate over(9), get(9), decide whether(9), ask(9), graduate(9), train(8), run on(8), compose of(8), be covered in(8), bring back(8), be called(8), offer(8), test(8), be dusted with(8), be selected by(8), gleam like(8), suck in(8), be among(8), see(8), transport(8), compete against(8), build with(8), assess(8), deal with(8), be sold in(8), say(8), be represented by(7), feature(7), be replaced by(7), be filled with(7), hold(7), deliver(7), read(7), dismiss(7), keep(7), be wrapped in(7), question(7), catapult(7), injure(7), direct(7), eliminate(7), recommend whether(7), . . .

0.29 NOMINALIZATION:ACT

Humans: involve(79), be made by(64), come from(36), affect(25), be done by(24), occur in(18), concern(16), be performed by(15), happen to(15), be caused by(15), include(14), happen in(12), remove(12), prevent(11), be initiated by(10), stimulate(9), harm(9), require(8), emerge from(8), eliminate(8), go into(8), explain(8), stop(8), regulate(7), interpret(7), split(7), be used by(7), be started by(7), be done to(7), study(7), reproduce(7), hurt(6), be in(6), divide(6), be related to(6), be about(5), help(5), limit(5), impact(5), be taken by(5), be of(5), originate from(5), examine(5), multiply(5), have(4), restore(4), create(4), be for(4), consist of(4), utilize(4), be undertaken by(4), recover(4), be from(4), enter(4), occur with(4), generate from(4), revive(4), restart(4), take back(4), use(3), happen with(3), be made about(3), produce(3), make(3), be made for(3), resemble(3), be done on(3), result from(3), translate(3), pertain to(3), service(3), discuss(3), erase(3), describe(3), be located in(3), center around(3), be by(3), occur to(3), relate to(3), be applied to(3), control(2), discourage(2), afflict(2), treat(2), be(2), start with(2), increase(2), save(2), be perpetrated on(2), engage(2), be given by(2), target(2), be aimed at(2), result in(2), inhibit(2), damage(2), be perpetrated by(2), destroy(2), delete(2), provide(2), reduce(2), occur within(2), deal with(2), support(2), effect(2), hold(2), look like(2), feature(2), belong to(2), accommodate(2), be tried by(2), be accomplished by(2), palpate(2), reclaim(2), accept(2), represent(2), be between(2), start(2), recreate(2), be blamed on(2), be used for(2), map(2), travel through(2), be conducted by(2), be created by(2), receive(2), bring about(2), investigate(2), explore(2), wipe out(2), avoid(2), be instigated by(2), be brought on by(2), analyze(2), harbor(2), tell about(2), be meant for(2), recapture(2), go through(2), be performed on(2), stem from(2), renew(2), pass through(2), soothe(2), start from(2), be done in(2), employ(2), manipulate(2), be made of(2), be generated by(2), ...

Web: affect(711), produce(173), allow(104), give(95), result in(78), be(67), occur in(55), cause(54), involve(51), generate(51), require(47), support(45), place(45), bring(41), benefit(41), be made by(40), make(38), divide(38), put(38), leave(37), harm(32), enable(30), drive(29), have(27), turn(25), teach(24), include(22), impact(21), prevent(20), limit(20), remove(20), be inflicted upon(20), form(18), provide(18), consider(17), hurt(16), help(16), permit(16), occur within(15), be perpetrated on(14), destroy(14), expose(14), separate(13), deprive(13), treat(12), assist(12), be inflicted on(12), be witnessed by(12), abuse(12), create(11), commit(11), scar(11), reflect(11), be associated with(11), protect(11), promote(11), delete(10), prohibit(10), force(10), live with(10), acne(10), convert(10), ongoing in(10), take(9), be made for(9), encourage(9), exclude(9), lead(9), be heaped upon(9), render(9), occur after(9), find(9), be observed in(9), say(9), whip(9), be made about(8), be for(8), uphold(8), segregate(8), undergo(8), be perpetrated against(8), be confirmed by(8), admit(8), keep(8), improve(8), withdraw(8), shape(8), connect(8), occur at(8), be in(7), be heaped on(7), be found in(7), touch(7), damage(7), come from(7), serve(7), kill(7), refer(7), be held after(7), yield(7), be alleged in(7), send(7), confront(7), invalidate(7), be upheld by(7), be correlated with(7), discourage(6), endanger(6), contain(6), effect(6), injure(6), indicate that(6), enhance(6), be presumed in(6), be followed by(6), influence(6), reflect on(6), be shaped by(6), parent(6), be reported in(6), be faced by(6), be expressed by(6), prepare(6), graduate(6), find that(6), be based on(6), characterize(6), compel(5), replace(5), hinder(5), concern(5), talk about(5), look at(5), mean that(5), befall(5), be perceived by(5), do(5), own(5), get(5), be accompanied by(5), occur before(5), attract(5), be compared with(5), say that(5), nourish(5), be made in(5), be visited upon(5), introduce(5), authorize(5), upset(5), ...

0.28 ABOUT

Humans: be about(133), involve(130), concern(79), discuss(41), be caused by(36), describe(31), relate to(26), talk about(25), deal with(22), refer to(22), come from(21), be related to(21), include(19), revolve around(16), contain(16), pertain to(16), focus on(14), result from(13), explain(13), cover(12), affect(12), feature(10), be over(10), address(9), reference(8), emerge from(7), tell about(7), center on(7), be concerned with(7), stem from(7), be due to(7), have(6), require(6), govern(6), be created by(6), portray(6), occur in(5), be(5), determine(5), regulate(5), detail(5), center around(5), teach(5), regard(5), highlight(5), happen because of(5), be regarding(5), occur with(4), encompass(4), be written about(4), tell of(4), relate(4), speak of(4), be on(4), show(4), use(4), be based on(4), surround(4), be inspired by(4), write about(4), define(4), be found in(4), encourage(3), incorporate(3), demand(3), promote(3), need(3), depict(3), be of(3), allow(3), complicate(3), support(3), look at(3), create(3), be against(3), argue(3), debate(3), derive from(3), question(3), teach about(3), be from(3), depend upon(2), report on(2), lose(2), legalize(2), examine(2), demonstrate(2), be in(2), present(2), review(2), approve(2), prohibit(2), study(2), praise(2), cause(2), mention(2), express(2), ban(2), comprise(2), effect(2), concentrate on(2), speak about(2), arise from(2), enforce(2), be for(2), depend on(2), be brought about by(2), be fought over(2), clarify(2), terminate in(2), generate from(2), analyze(2), happen due to(2), go into(2), be because of(2), photograph(2), empower(2), decide(2), be caused for(2), overuse(2), be comprised of(2), employ(2), talk of(2), occur because of(2), be precipitated by(2), elaborate on(2), decide on(2), threaten(2), argue about(2), expose(2), investigate(1), remind of(1), explore(1), embody(1), celebrate(1), prescribe(1), exude(1), be decided by(1), call for(1), arise during(1), mourn(1), control(1), set(1), lay out(1), be filled with(1), prevent(1), provide(1), mandate(1), tell(1), be classified as(1), . . .

Web: impose(476), be(273), make(171), affect(168), follow(147), involve(136), include(107), cover(103), raise(95), reduce(94), have(93), deal with(78), feature(75), be about(74), focus on(71), tell of(68), do(67), describe(64), pay(62), levy(61), celebrate(54), require(51), relate(51), recount(49), express(48), be covered by(44), increase(42), combine(42), tell(35), allow(34), speak of(34), chronicle(34), establish(32), go(32), continue(31), cut(31), place(31), explore(30), reflect(27), lower(27), go beyond(27), create(26), say(25), provide for(24), get(23), concern(22), cause(22), govern(21), tell about(21), put(21), revolve around(20), demand(20), mean(19), talk about(19), depict(18), show(18), present(17), reveal(17), regulate(16), teach(16), arise from(16), provide(16), bring(16), compare(16), bear upon(16), address(15), detail(15), prohibit(15), result in(15), give(15), make up(15), introduce(15), depend upon(14), promote(14), look at(14), eliminate(14), be collected in(14), be known as(14), need(13), offer(13), limit(13), set(12), collect(12), contain(12), write(12), be filled with(11), specialize in(11), represent(11), influence(11), call(11), charge(11), conflict with(11), become(11), be in(10), result from(10), consider(10), examine(10), impact(10), define(10), inspire(10), appear in(10), embody(9), be influenced by(9), equalize(9), draw from(9), reflect on(9), be addressed in(9), send(9), differ from(9), sing i(9), prevent(8), narrate(8), go down in(8), be within(8), go in(8), permit(8), capture(8), trace(8), seraphs(8), accompany(8), phase out(8), smell with(8), be told of(8), avoid(8), focus upon(8), epitomize(8), generate(8), lads(8), encompass(7), be based on(7), ban(7), regard(7), enforce(7), fall in(7), violate(7), push(7), distinguish between(7), convey(7), promise(7), be associated with(7), envisage(7), say that(7), repeal(7), originate in(7), supersede(7), authorize(7), encourage(6), emerge from(6), be classified as(6), be of(6), . . .

0.12 HAVE$_2$

Humans: belong to(85), come from(72), be owned by(54), surround(37), protect(24), be possessed by(19), cover(18), be in(15), involve(15), be on(13), be used by(12), be found on(11), be held by(10), enclose(10), be of(9), encircle(9), grow on(7), be found in(7), support(7), weigh(6), encompass(6), emerge from(6), be inside(6), concern(6), attend(5), be within(5), be derived from(5), be controlled by(5), be managed by(5), be created by(5), be generated by(5), affect(5), be passed down through(5), define(5), happen to(5), describe(5), be related to(5), be had by(5), measure(4), join(4), be located in(4), hold(4), be made by(4), be associated with(4), be from(4), be experienced by(4), come off(4), smell like(4), go to(4), relate to(4), include(3), be taken from(3), be produced by(3), derive from(3), be caused by(3), reside in(3), empower(3), make up(3), be exercised by(3), border(3), originate from(3), lie within(3), be faced by(3), be about(3), pass through(3), be given to(3), be demonstrated by(3), be maintained by(3), be attributed to(3), refer to(3), occur with(2), have(2), go around(2), be kept by(2), be administered by(2), be at(2), circle(2), be kept in(2), result from(2), worry(2), represent(2), work for(2), be employed by(2), fill(2), bother(2), plague(2), evaluate(2), occur within(2), reach(2), be around(2), envelop(2), emanate from(2), form(2), comprise(2), originate with(2), contain(2), enable(2), vote for(2), fit(2), supply(2), run through(2), participate in(2), arise from(2), appear on(2), be for(2), stay in(2), be seized by(2), be involved with(2), found within(2), descend in(2), be wielded by(2), be made for(2), be posessed by(2), be shown by(2), be inside of(2), be reached by(2), be linked to(2), be surrounded by(2), grow from(2), coat(2), be exerted by(2), shield(2), be made from(2), encourage(1), hurt(1), balance(1), carry(1), be sold by(1), motivate(1), come with(1), be taken by(1), frustrate(1), emerge in(1), help(1), mark(1), guard(1), separate(1), move(1), control(1), harm(1), constitute(1), compose(1), block(1), accept(1), sit in(1), drain(1), reside within(1), isolate(1), afflict(1), be connected with(1), ...

Web: be(3805), surround(709), require(598), affect(529), enable(388), allow(385), be owned by(287), leave(237), represent(230), protect(218), confront(217), prevent(204), encircle(203), divide(185), join(172), cause(170), lead(161), make(139), ask(120), be in(117), have(113), be acquired by(111), help(109), keep(106), encourage(103), interfere with(99), bring(96), engage(93), force(86), arise between(82), be held by(80), become(79), support(76), be used by(76), hold(75), attend(74), enclose(72), contract(72), challenge(72), face(72), attract(67), serve(62), separate(58), include(58), be adopted by(58), take(55), provide(54), involve(52), impact(52), motivate(51), want(50), switch(48), oppress(47), give(47), turn(47), drive(45), draw(44), be purchased by(41), ring(41), assist(41), be solved by(40), plague(40), increase(40), deprive(40), arise from(39), be taken by(39), change(39), come from(38), kill(37), form(36), prepare(36), be granted by(36), defend(35), fit(35), put(35), meet(35), quit(35), arise for(35), carry(34), be of(34), arise with(34), expose(33), test(33), impede(32), benefit(32), be expelled from(32), arise in(31), hinder(30), transform(29), be selected by(28), resign from(28), be affiliated with(28), stand between(27), book(27), threaten(27), be used with(27), move(26), encompass(26), be leased from(26), mean that(26), destroy(26), be sold by(25), oppose(25), embarrass(25), be vested in(24), be built around(23), be acquired from(23), set(23), appear on(22), influence(22), perplex(22), be that(22), remain in(22), control(21), be amassed by(21), exist in(21), exploit(21), arise during(21), defect from(21), reflect(21), be confiscated by(21), permit(21), be with(21), prompt(21), guard(20), strengthen(20), save(20), be registered with(20), disagree with(20), prohibit(20), hurt(19), result from(19), run through(19), define(19), be included in(19), be found on(19), limit(19), be given by(19), impair(19), connect(19), be delegated by(19), be set by(18), introduce(18), sway(18), push(18), be appointed by(18), need(18), enhance(18), be made(18), be embedded in(18), ...

E.4 Comparison by Class: Using The First Verb Only

0.97 CAUSE$_1$

Humans: cause(91), promote(11), create(10), contain(7), have(5), make(5), induce(5), produce(5), carry(4), evoke(4), generate(4), result in(3), be(3), stimulate(3), elicit(3), spread(2), regulate(2), lead to(2), build(1), precipitate(1), feed(1), provide(1), involve(1), provoke(1), give(1), include(1), help in(1), be filled with(1), affect(1), be condusive to(1), s purpose be(1), make for(1), supply(1), happen by(1), be produced by(1), foster(1), be found in(1), be made of(1), form(1), spark(1), supplement(1), percipitates(1), illicits(1), occur by(1), come from(1), be conducive to(1), cause to(1), shrades(1), aid(1), be about(1).

Web: cause(1949), stimulate(403), regulate(222), promote(211), produce(109), affect(105), control(67), be(52), include(48), spread(44), have(42), inhibit(31), influence(29), rely on(26), create(25), induce(25), carry(24), give(24), encourage(24), build(22), trigger(22), make(21), accelerate(17), differentiate(17), stunt(16), combine(14), result in(13), spur(12), modulate(12), bring(11), govern(11), be required for(10), be like(10), limit(10), propagate(9), prevent(9), use(9), determine(9), stop(9), support(9), be associated with(9), suppress(8), lack(8), work with(8), help with(8), be involved in(8), choke as(8), miss(8), be filled with(7), mimic(7), be needed for(7), fuel(7), offer(7), drive(7), ensure(7), transmit(7), display(6), slow(6), be trapped in(6), act as(6), work as(6), become(6), generate(5), modify(5), maintain(5), initiate(5), increase(5), blend(5), be for(5), provide(4), contain(4), develop from(4), help(4), twist around(4), activate(4), curb(4), be used in(4), pass on(4), produce that(4), reduce(4), impair(4), tamponades(4), be confirmed by(4), capture(4), be implicated in(4), replace(4), deliver(4), build on(4), imply(4), stretch(4), allow for(4), suggest that(4), bring about(4), provoke(3), help in(3), balance(3), improve(3), be mediated by(3), call(3), breast(3), isn(3), mine(3), be characterized by(3), contract(3), add(3), start(3), build in(3), do(3), cause that(3), utilize(3), sow(3), be concealed by(3), facilitate(3), understand(3), ignite(3), enable(3), range from(3), be made from(3), spread like(3), depend on(3), be unlike(3), feed off(3), feed(2), involve(2), be unleashed after(2), be focused on(2), speed(2), retard(2), rack up(2), plod along with(2), result from(2), pass(2), advance(2), resemble(2), shoot(2), be called(2), exist during(2), veer towards(2), aid in(2), appear as(2), express(2), start from(2), mix(2), pile on(2), be released include(2), arrest(2), livestock(2), disappear as(2), break(2), suffer(2), ...

0.89 MAKE$_1$

Humans: produce(23), make(22), sing(19), play(15), have(4), create(3), give(2), supply(2), deliver(1), include(1), manufacture(1), contain(1), exude(1), provide(1), warble(1), store(1), ooze(1), be made for(1), ooze out(1), grow(1), love(1), erupt in(1).

Web: make(333), produce(295), sing(264), have(196), play(105), gather(104), learn(80), collect(57), force(52), be(45), hear(41), suck(40), give(35), exude(33), drip(33), provide(30), carry(26), get(26), bring(24), contain(20), ooze(20), draw(20), lose(20), supply(19), extract(19), cause(18), continue in(18), yield(18), create(17), know(17), drop(16), drink(16), secrete(16), allow(15), accumulate(13), be tapped for(12), raid(11), develop(11), bleed(11), store(10), eat(10), be kept for(10), be in(10), mimic(10), do(10), acquire(10), be attracted by(10), leak(9), share(9), change(9), vary(9), look for(9), find(9), seek(9), turn(9), look like(9), rely on(9), break into(8), access(8), resemble(8), consume(8), close(8), include(7), prevent(7), add(7), exhaust(7), copy(7), emit(7), feed on(7), burst with(7), sting(7), say(7), deliver(6), live on(6), be without(6), leave(6), be known as(6), forget(6), offer(6), complete(6), feature in(6), feature(6), flow(5), handle(5), master(5), prepare(5), modify(5), move(5), fill(5), simulate(5), steal(5), defend(5), fly like(5), send(5), sip(5), take(5), weep(5), amplify(5), bring in(5), hold(5), degrade(5), hive(5), warble(4), transport(4), lack(4), help(4), blast(4), taperecorded(4), call(4), feed(4), be called(4), send out(4), promise(4), generate(4), wanteth(4), be harvested for(4), run on(4), reach(4), whisper(4), pick up(4), reproduce(4), store up(4), stream(4), be checked for(4), utter(4), trill(4), roam for(4), use be(4), be cut off that(4), knead up(4), spread from(3), twitter(3), taste(3), come with(3), begin(3), lay up(3), improve(3), be on(3), represent(3), carry over into(3), be known for(3), combine(3), deposit(3), instill(3), use(3), discriminate between(3), possess(3), obtain(3), want(3), enjoy(3), ...

0.87 BE

Humans: be(231), resemble(86), look like(68), involve(46), act like(31), include(21), act as(18), be like(17), come from(17), use(13), contain(13), have(12), be shaped like(12), consist of(11), be composed of(11), be made of(10), be called(9), behave like(8), work as(8), cover like(8), become(7), be related to(7), be made up of(6), relate to(6), mimic(5), require(5), be classified as(5), emulate(5), hit(5), be about(5), be at(4), concern(4), represent(4), describe(4), grow into(4), live like(4), belong in(4), be made from(4), recall(4), swim like(4), refer to(4), be in(3), feel like(3), smell like(3), be based on(3), serve as(3), be comprised of(3), form(3), work like(3), be concerned with(3), be for(3), be considered(3), be of(3), be named(3), be birthed by(3), be found in(3), pertain to(3), be used as(2), revolve around(2), occur in(2), incorporate(2), be sold as(2), emerge from(2), be run by(2), smell of(2), exist as(2), reach(2), bear(2), occur at(2), be made by(2), occur during(2), produce(2), come in(2), accompany(2), depend on(2), rely on(2), belong to(2), remind of(2), be used to(2), be born of(2), derive from(2), go in(2), exemplify(2), appear(2), center around(2), espouse(2), go to(2), flow in(2), be regarding(2), descend from(2), be had by(2), be aimed at(1), appear like(1), house(1), encompass(1), help(1), work with(1), display(1), seem(1), make(1), touch(1), protect(1), be formed by(1), move(1), operate(1), constitute(1), be employed as(1), accept(1), be filled with(1), invite(1), provide(1), go into(1), live as(1), duplicate(1), be characterized as(1), be regarded as(1), generate(1), do(1), determine(1), enjoy(1), challenge(1), like(1), exhibit(1), coincide with(1), play(1), wear(1), comprise(1), deal with(1), teach(1), cover(1), meet as(1), warm(1), love(1), lead(1), hold(1), aim at(1), seem like(1), supply(1), be treated like(1), specialize in(1), be from(1), work on(1), perform(1), recruit(1), be described as(1), claim(1), portray(1), be created out of(1), be thin like(1), be curved like(1), ...

Web: be(3949), include(955), help(397), work with(276), provide(233), support(228), have(218), become(218), assist(210), look like(190), serve(167), resemble(156), represent(137), play(132), work as(127), bring(123), involve(112), make(95), serve as(95), range from(93), teach(91), educate(81), allow(75), form(72), enable(72), encourage(70), be at(67), recognise(62), reach(56), give(49), capture(48), study(48), be called(46), be considered(44), train(44), promote(42), consist of(40), use(39), require(38), connect(38), know(37), appear as(36), win(35), attempt(34), act as(33), lose(33), be in(32), be created by(32), portray(31), put(31), exceed(31), compete with(30), find(30), warm(28), specialize in(28), empower(28), match(28), follow(28), be like(27), occur in(26), want(26), take(26), be founded by(26), deal with(24), launch(24), link(24), reduce(24), live as(23), offer(22), host(22), feature(22), contain(21), be affiliated with(21), make up(21), organize(21), function as(21), recognize(21), constitute(20), be treated as(20), meet as(19), kill(19), join(19), lie at(19), guide(19), be run by(18), write(18), be brought up with(18), start as(18), prevent(18), get(18), cut(18), tell(18), surround(18), be comprised of(17), be during(17), inspire(17), call(17), lead(16), accompany(16), comprise(16), hold(16), shoot(16), reflect(16), affect(16), occur as(16), be formed by(15), touch(15), rely on(15), treat(15), be associated with(15), benefit(15), visit(15), coincide with(14), love(14), save(14), marry(14), advise(14), care about(14), engage(14), be started by(14), mimic(13), be classified as(13), mean that(13), advocate for(13), drive(13), inform(13), believe that(13), be used as(12), incorporate(12), meet with(12), increase(12), pose as(12), focus on(12), drive with(12), introduce(12), think(12), modify(12), sponsor(12), do(11), generate(11), think that(11), be represented as(11), mobilize(11), run from(11), surpass(11), establish(11), name(11), dehumanize(11), murder(11), expect(11), ...

0.65 CAUSE$_2$

Humans: be caused by(168), come from(59), result from(22), be made by(20), be created by(12), involve(7), be from(7), be due to(6), occur in(4), contain(4), occur during(4), be induced by(3), occur from(3), happen during(3), happen in(3), be eaten by(3), be generated by(3), occur because of(3), be caused from(2), cause(2), indicate(2), look like(2), accompany(2), control(2), occur at(2), be left by(2), be made from(2), happen because of(2), occur with(1), have(1), occur on(1), result in(1), coincide with(1), precede(1), be affected by(1), force(1), affect(1), resemble(1), come before(1), be produced by(1), be associated with(1), be exerted by(1), show(1), stem from(1), be blamed on(1), use(1), arise from(1), recognize(1), react to(1), occur within(1), be done with(1), be linked to(1), be related to(1), come froom(1), be frayed from(1), be measured in(1), emerge from(1), run on(1), stain(1), be relative to(1), emanate(1), be provoked by(1), flow with(1), be measured from(1), be exhibited by(1), be made with(1), protect(1), remain(1), emanate from(1), bring on by(1), develop in(1), be brought on by(1), be found in(1), form(1), be made of(1), stain like(1), be produced from(1), happen with(1), be dismayed about(1), be eaten out by(1), concern(1), be filled with(1), effect(1), be had during(1), need(1), be created from(1), be jangled from(1), emboss(1), derive from(1), relate to(1), invade(1), be derived from(1), emit from(1), occur do to(1), be cuased by(1), be gotten from(1), appear with(1), occur due to(1), delineate(1), exemplify(1), refer to(1), seal in(1), show as(1), complicate(1), utilize(1), arise due to(1), be cause by(1), be about(1).

Web: be caused by(290), result from(217), cause(160), draw(109), be(97), be characterized by(81), have(76), be associated with(68), force(67), accompany(56), give(56), involve(54), result in(43), pull(41), arise from(39), prevent(34), drive(23), come with(22), make(19), suck(19), contaminate(18), bring(17), occur in(16), precede(16), come from(16), be covered by(16), follow(16), keep(15), push(15), attend(14), include(14), allow(14), deliver(14), blow(13), indicate(12), draw in(12), occur from(10), be exerted by(10), be suspended after(10), affect(8), move(8), stand for(8), reflect(8), supply(8), be removed with(8), cool(7), be fueled by(7), reveal(7), squeeze(7), occur with(6), move through(6), be removed by(6), occur without(6), be called(6), mention(6), be reversed by(6), create(6), exist in(6), mean(6), be governed by(6), be separated from(6), depend on(6), resemble(5), show(5), be blamed on(5), compress(5), indicate that(5), be inundated by(5), modify(5), herald(5), mean that(5), rise(5), accelerate(5), be diagnosed as(5), persist in(5), be characterised by(5), melt into(5), stem from(4), use(4), be from(4), look like(4), depend upon(4), develop from(4), go with(4), set(4), be maintained in(4), require(4), pressurize(4), represent(4), propagate through(4), provide(4), attract(4), be observed with(4), condition(4), occur(4), be aggravated by(4), be caused(4), chop(4), retain(4), be correlated with(4), displace(4), be related with(4), induce(4), be induced by(3), occur on(3), ascend in(3), manifest with(3), get into(3), start as(3), drag(3), be hastened by(3), reverberate into(3), let(3), be on(3), leak(3), denser(3), wash away with(3), live in(3), be endured before(3), be exacerbated by(3), be marked by(3), stop(3), be accompanied by(3), remove(3), be caused be(3), surround(3), prompt(3), take(3), be recognized as(3), prepare for(3), develop between(3), go away with(3), extend over(3), posit that(3), be washed out with(3), entrain(3), drag in(3), be affected by(2), coincide with(2), be made by(2), recognize(2), divert(2), exist between(2), score(2), push inside(2), consist in(2), be supplied through(2), direct(2), occur after(2), be obliterated by(2), ...

0.63 NOMINALIZATION:AGENT

Humans: give(25), avoid(18), cut(14), sort(12), plan(9), work with(6), donate(5), slice(5), design(4), take(4), work for(3), organize(3), process(3), come from(2), lay out(2), count(2), handle(2), ignore(2), run away from(2), work on(2), edit(2), evade(1), supply(1), oversee(1), study(1), divide(1), dodge(1), deal with(1), collect(1), write(1), hide from(1), tend to(1), trim(1), collect data for(1), run through(1), participate in(1), pertain to(1), enumerate(1), brainstorm(1), live in(1), measure(1), make(1), be used on(1), splice(1), run from(1), administer(1), document(1), look through(1), conduct(1), log(1), be used for(1), canvas for(1), complete(1), arrange(1), subvert(1), monitor(1).

Web: give(654), donate(395), receive(74), sell(41), provide(40), work with(27), sort(25), work for(25), serve on(22), cut(20), supply(17), be(15), work in(14), design(13), treat(12), build(11), match(11), contribute(10), have(9), offer(9), work at(9), mean(9), see(9), donate whole(8), visit(8), dodge(7), deal with(7), include(7), understand(7), believe that(7), evade(6), lay out(6), write(6), look at(6), create(6), view(6), manage(6), divide(5), plan(5), study(5), restructure(5), conceive of(5), develop(5), choose(5), consider(5), give whole(5), transform(5), craft(5), giv(5), meet(4), coordinate with(4), saw(4), tour(4), love(4), represent(4), promote(4), exclude(4), be with(4), sweep(4), work between(4), qualify as(4), be employed by(4), serve(4), be transmitted by(4), want(4), coordinate(4), take(3), advise(3), value(3), head(3), remake(3), fight for(3), foster(3), conform(3), deplore(3), favor(3), think that(3), can-donate(3), separate(3), believe(3), destroy(3), chop up(3), envision(3), be vaccinated for(3), be given(3), be of(3), be told(3), consult with(3), be hired by(3), come from(2), be recruited by(2), occur in(2), care for(2), ruin(2), incorporate(2), identify(2), protest(2), perforate(2), be exposed from(2), lobby(2), be housed within(2), be located within(2), transform of(2), argue that(2), realize that(2), assist(2), designate(2), be infected with(2), take over(2), carve(2), be involved in(2), staff(2), pass(2), sever(2), pick(2), be spiked with(2), mean that(2), know(2), need(2), tell(2), urge(2), rewrite(2), share(2), be employed in(2), defy(2), enter(2), selflessly(2), serve as(2), allow(2), criticize(2), oversee(1), avoid(1), collect(1), join(1), book(1), draw(1), estimate that(1), act in(1), indicate that(1), result in(1), forget(1), prosecute(1), help(1), spare(1), disclose(1), operate in(1), experiment with(1), write on(1), act as(1), give of(1), . . .

0.58 USE

Humans: use(70), be done by(14), come from(13), be made of(13), utilize(11), be made by(11), involve(10), be operated by(9), require(9), be made from(9), produce(7), rely on(6), cook with(6), emit(5), run on(5), cause(5), be(5), be worn on(5), work with(4), be used by(4), move(4), need(4), be performed by(4), be caused by(4), be played by(4), be applied by(3), contain(3), operate by(3), induce(3), be played with(3), emerge from(3), be composed of(3), consist of(3), employ(2), include(2), burn(2), mimic(2), go on(2), be powered by(2), be like(2), work on(2), be related to(2), be comprised of(2), attach to(2), emanate from(2), generate from(2), clean by(2), lead to(2), cook under(2), occur through(2), be given by(2), be based on(2), turn use(2), be transmitted over(1), be broadcast by(1), incorporate(1), be taken by(1), consume(1), make(1), work in(1), come through(1), be by(1), be performed in(1), turn with(1), ignite(1), spin(1), be fueled by(1), be generated by(1), work by(1), resemble(1), provide(1), function as(1), be worked by(1), release(1), go in(1), simulate(1), generate(1), do(1), look like(1), allow(1), depend on(1), rely upon(1), flow through(1), count(1), occur in(1), be operated with(1), be expressed by(1), be conducted by(1), transport(1), wear on(1), happen by(1), give off(1), determine by(1), be applyed by(1), result in(1), flow with(1), heat by(1), cook use(1), be used on(1), vent(1), happen in(1), be made with(1), signal(1), administer(1), help in(1), exumes(1), suck in(1), invloves(1), operate use(1), consider(1), happen over(1), pick up with(1), be cast using(1), run by(1), involve use(1), be pushed with(1), occur via(1), take out(1), concern(1), be based upon(1), occur because of(1), host(1), be specific for(1), happen from(1), build up(1), be counted by(1), create(1), transmit to(1), be ascribed to(1), tumble(1), puff(1), suck by(1), pertain to(1), be changed by(1), be done over(1), practice(1), be transmitted through(1), feel like(1), change(1), turn from(1), be transmitted by(1), turn by(1), be developed by(1), reside in(1), steam(1), cook by(1), be rooted in(1), work via(1), describe(1), suck into(1), . . .

Web: be(160), use(150), be operated by(112), provide(47), include(46), receive(44), raise(44), run on(38), burn(33), lift(32), pump(30), have(28), bring(28), make(27), be turned by(26), produce(24), draw(24), involve(22), spill(20), allow(18), generate(17), be broadcast by(14), be applied by(14), require(13), be powered by(13), drive(12), invent(12), give(11), deliver(11), be transmitted over(10), prevent(10), be heated by(10), cause(9), represent(9), operate at(9), convert(9), churn(9), be like(8), own(8), maintain(8), seem like(8), work with(7), need(7), rely on(7), run over(7), take(7), be used with(7), discharge(7), be wrought by(7), move(6), consume(6), contain(6), look like(6), carry(6), leak(6), absorb(6), be controlled by(6), be under(6), operate on(6), scoop(6), reduce(6), develop(6), supply(6), be run with(6), appear on(6), work in(5), be fueled by(5), do(5), utilize(5), emit(5), resemble(5), employ(5), be used as(5), mean that(5), be produced by(5), be called(5), be released by(5), dominate(5), occur over(5), tend toward(5), be rejected by(5), pressurize(5), establish(5), smell of(5), pick up than(5), be set by(5), sell(5), squeeze(5), purify(5), empty(5), be driven by(5), remove(5), enable(5), simulate(4), be performed by(4), be by(4), draw up(4), be accomplished by(4), be corrected by(4), control(4), lift up(4), convey(4), design(4), combine(4), comprise(4), rest in(4), run(4), mean(4), be activated by(4), accompany(4), turn off(4), admit(4), fuse(4), detonate(4), be filled with(4), be behind(4), require that(4), work like(4), turn(4), mimic(3), turn with(3), be generated by(3), be worked by(3), depend on(3), spray(3), throw(3), be mounted on(3), advocate(3), ignore(3), pass by(3), operate(3), be obtained by(3), apply(3), cook(3), slosh(3), be designed for(3), be influenced by(3), secure(3), splash(3), be known by(3), notify(3), be of(3), be imported in(3), live(3), permit(3), spit(3), be propelled in(3), understand(3), don(3), muffle(3), . . .

0.51 HAVE$_1$

Humans: contain(87), have(52), be made of(34), be made from(25), bear(18), come from(11), be made with(8), produce(8), be composed of(7), include(5), use(5), own(5), consist of(5), display in(5), be in(4), display(3), be made using(3), nourish(3), harbor(3), utilize(3), provide(2), be garnished with(2), show(2), taste like(2), be made up of(2), love(2), be inhabited by(2), be filled with(2), project in(2), be comprised of(2), show in(2), care for(1), host(1), resemble(1), bare(1), feed(1), supply(1), add(1), adopt(1), like(1), have within(1), keep(1), be like(1), employ(1), be trimmed with(1), be decorated with(1), form(1), grow(1), be(1), look like(1), allow(1), support(1), rely on(1), enjoy(1), feature(1), caintains(1), be home to(1), project with(1), pull out(1), comtains(1), be attached to(1), encompass(1), accomodates(1), bloom(1), be formed from(1), begin because of(1), be baked with(1), be able to make(1), originate from(1), transmit in(1), be shaped as(1), be of(1), perform(1), be flavored with(1), be populated with(1), possess(1), be made out of(1), nurture(1), be trimmed in(1), broadcast in(1), illuminate(1), be high in(1), be thickened with(1), be cooked with(1), be feels like(1), sorround(1).

Web: bear(2114), have(1885), produce(1230), contain(441), include(435), give(428), be(383), bore(269), provide(222), yield(203), show(201), bring(194), bringeth(140), paint(119), use(89), lose(87), adopt(85), grow(76), feature(68), present(68), carry(67), drop(53), offer(53), be loaded with(47), love(46), want(46), be filled with(43), own(38), keep(38), set(37), do(36), combine(32), take(31), beareth(27), be made with(26), consist of(26), be in(25), make(23), hold(22), get(21), need(20), lack(19), be illustrated with(18), fill(17), supply(15), draw(15), describe(15), give up(15), taste like(14), bare(14), look for(14), have in(14), become(14), allow(13), know(13), put(13), house(13), form(12), protect(12), travel with(12), introduce(12), attend(11), furnish(11), put forth(11), be called(11), cause(11), cast(10), be than(10), look at(10), be made of(9), rely on(9), range from(9), afford(9), like(8), be like(8), be used as(8), be laden with(8), treat(8), leave(8), take in(8), be packed with(8), consider(8), collect(8), preserve(8), be located within(8), create(8), dissect(8), shed(8), miss(8), care for(7), host(7), be composed of(7), support(7), be made from(7), abandon(7), throw(7), call for(7), bring forth(7), be used in(7), puree(7), be grown for(7), be harvested for(7), retain(7), enjoy(6), hurt(6), sprout(6), teach with(6), rotten(6), dry(6), register(6), maintain(6), be known as(6), be based on(6), be mixed with(6), find(6), grow from(6), look after(6), display(5), feed(5), add(5), look like(5), foster(5), move(5), match(5), care about(5), inspire(5), be used for(5), be ornamented with(5), live(5), start with(5), interpret(5), raise(5), avoid(5), be on(5), dissolve(5), appear in(5), be cultivated for(5), hang(5), colour(5), be enriched with(5), go(5), be stripped of(5), invent(5), resemble(4), incorporate(4), maketh(4), substitute(4), be seen as(4), encounter(4), take on(4), ...

0.50 FROM

Humans: come from(170), be made from(90), be made of(44), be made in(16), live on(14), be from(12), be conceived in(12), contain(10), be produced from(9), be made at(9), live in(8), work on(6), be bought in(6), be made with(5), be used in(5), be derived from(5), be(5), go to(5), be raised in(4), be found in(4), be born in(4), visit(4), emerge from(4), be created in(4), derive from(4), be processed from(4), work in(3), be distilled from(3), resemble(3), originate from(3), be rendered from(3), be composed of(3), be created from(3), be bought from(2), be produced in(2), be extracted from(2), belong in(2), be sold in(2), remind of(2), come out of(2), be purchased in(2), be produced by(2), originate at(2), be bred in(2), stick to(2), make of(2), reside on(1), make(1), buy(1), be taken from(1), be done at(1), be obtained from(1), be in(1), be used at(1), hail from(1), be kept in(1), be on(1), arrive from(1), represent(1), smell of(1), come off(1), make from(1), blow off(1), taste like(1), use(1), visit from(1), be pressed from(1), be generated from(1), be found at(1), belong to(1), be provided at(1), be created with(1), be attached to(1), help at(1), be formulated from(1), incorporate(1), be bought at(1), taste(1), be comprised of(1), be fermented from(1), be located in(1), be bought(1), be born of(1), be heard in(1), be born from(1), in make from(1), be formed by(1), come off of(1), holiday in(1), be made using(1), be pressed out of(1), be exuded by(1), be made out of(1), be created by(1), be used with(1), be spread by(1), consist of(1), result from(1), be fried out of(1), be trimmed from(1), be carved off(1), relate to(1), originate in(1), be formed from(1), pertain to(1), be near(1), go in(1), be carved from(1), stem from(1), travel to(1), come to(1), be made up of(1), be compose of(1), be purchased from(1), be skin from(1), be fertilized in(1).

Web: come from(204), live on(175), be(146), enter(134), be in(113), pee in(110), leave(107), grow up on(103), work on(72), blow from(72), have(57), contain(53), visit(49), be made from(42), include(38), come into(34), be from(32), live in(29), be sold in(29), resemble(26), arrive in(24), be extracted from(23), make(22), taste like(21), be sold as(21), arrive at(21), reside in(21), stay in(18), arrive from(17), be obtained from(16), be prepared at(16), come off(15), be produced from(15), cause(15), remain on(13), blow towards(13), ruffle(13), be substituted for(13), use(12), be kept in(12), be derived from(12), be produced in(12), be raised on(12), be raised in(11), be tried at(11), sell in(11), raise(11), go from(11), be pressed from(10), be made at(10), descend on(10), produce(10), remain in(10), work(10), be found in(9), frilled(9), offer(9), tour(9), hit(9), occur in(9), taste of(9), be reared on(9), love(9), be used in(8), ebb over(8), come over(8), travel in(8), rival(8), be produced with(8), sell(8), rise with(8), fall from(8), compete with(8), hail from(7), know(7), tame(7), be refined(7), be born on(7), rise from(7), work at(7), turn(7), be termed(7), represent(6), buy(6), be done at(6), gambol over(6), require(6), be manufactured from(6), show(6), take(6), replace(6), view(6), be sold at(6), be blended with(6), be sold by(6), be at(6), divide(6), flow from(6), look like(6), smell of(5), be made of(5), fit in(5), predominate over(5), be pressed(5), go into(5), be marketed as(5), fly from(5), run(5), write about(5), be associated with(5), be labeled(5), take over(5), give(5), fuel(5), inherit(5), hold(5), evoke(5), press(5), reside on(4), be distilled from(4), be used at(4), be on(4), be generated from(4), be born in(4), be rendered from(4), domesticate(4), build(4), be wafted across(4), fly over(4), stumble upon(4), travel through(4), be purchased at(4), be squeezed from(4), like(4), be packed with(4), travel between(4), consider(4), run from(4), be administered at(4), result in(4), compare with(4), be than(4), revitalize(4), stay on(4), smell like(4), pass through(4), be mixed with(4), be bought by(4), saw(4), be trained on(4), ...

0.42 FOR

Humans: contain(40), treat(20), be for(17), dispense(16), be used for(16), attract(14), cure(12), manufacture(12), be used in(11), make(10), try(8), hold(8), feed(8), be used on(7), nourish(7), produce(7), include(6), involve(6), have(5), specialize in(5), relieve(5), feature(5), come from(4), relate to(4), go in(4), heal(4), draw(3), handle(3), pump(3), sell(3), focus on(3), conduct(3), care for(3), channel(3), pay for(3), be made for(3), know about(3), study(3), house(2), prosecute(2), help(2), direct(2), deal with(2), aid(2), build(2), go into(2), vend(2), serve(2), do(2), process(2), tend to(2), arrest(2), be placed in(2), go on(2), cover(2), concern(2), be on(2), be worn by(2), distribute(2), work on(2), facilitate(1), encompass(1), assemble(1), spew(1), work with(1), work in(1), touch(1), buy(1), understand(1), be used by(1), be found in(1), offer(1), be in(1), consist of(1), deter(1), affect(1), be filled with(1), know(1), provide(1), create(1), adjudicate(1), specialise in(1), give(1), help with(1), benefit(1), reduce(1), investigate(1), harness(1), moisturize(1), practise on(1), be inserted into(1), be supplied by(1), drop into(1), be located in(1), nurishes(1), be given to(1), adjucates for(1), be eaten by(1), demonstrate(1), be exuded by(1), safeguard from(1), consider(1), drain from(1), underscore(1), clear(1), be involved in(1), deflect(1), be utilized for(1), be spent on(1), be designed for(1), deaden(1), be earmarked for(1), be meant for(1), go onto(1), be devoted to(1), govern(1), negate(1), work at(1), punish(1), invite(1), fund(1), be active for(1), be applied on(1), highlight(1), cater to(1), pertain to(1), provide for(1), deoderizes(1), pump up(1), be given(1), be allocated to(1), research(1), fence in(1), be applied to(1), protect from(1), be set aside for(1), be fed to(1), aid in(1), fix(1), be absorbed by(1), welcome(1), be intended for(1), deodorize(1), to play(1).

Web: produce(709), have(375), contain(349), include(260), be(167), come from(155), find(85), be derived from(74), give(63), test(58), be used for(49), pump(44), interrupt(43), need(43), cause(41), kill(38), hold(36), show(36), feed(34), flow(34), allow(34), be filled with(33), be drilled for(32), feature(31), make(30), keep(30), state that(29), sell(27), operate(25), yield(22), place(22), be used in(20), encounter(20), target(20), display(19), strike(19), commit(18), attract(17), consist of(16), handle(16), dispense(16), specialize in(16), discover(16), cure(15), deal with(15), hit(15), paint(14), be utilized in(14), build(13), nourish(13), adjudicate(13), be struck by(13), draw(12), deter(12), sit as(12), identify(12), push(11), use(11), don(11), work in(10), be used by(10), require(10), be designated as(10), do(9), offer(9), ride(9), prevent(9), be melted by(9), convey(9), own(9), drive(9), harm(9), present(9), be known as(9), confirm(9), develop(9), release(9), be completed as(9), take away(9), provide(8), be produced by(8), turn away(8), play(8), take(8), recover(8), travel by(8), receive(8), be from(8), see(8), be administered by(8), resemble(8), supply(8), be made from(8), treat(7), manufacture(7), bring(7), be ripened on(7), carry(7), be produced in(7), allow for(7), look like(7), be in(6), try(6), serve(6), tap(6), leak(6), stay within(6), work for(6), run down(6), combine(6), capture(6), drill for(6), be provided by(6), put(6), modify(6), be obtained from(6), prove(6), stop(6), enable(6), work with(5), relieve(5), house(5), assemble(5), dislike(5), purchase(5), pump out(5), be suited for(5), apply throughout(5), comprise(5), be made by(5), reverse(5), be associated with(5), hear(5), go with(5), produce as(5), reveal(5), grip(5), adjudge(5), turn(5), transmit(5), sustain(5), be used on(4), reduce(4), be for(4), serve in(4), come with(4), insure(4), force(4), be established by(4), spout(4), ...

0.37 IN

Humans: live in(58), occur in(39), be in(35), occur during(23), occur at(19), happen in(16), be eaten at(15), occur on(13), involve(13), be located in(12), be played in(11), happen at(10), come from(9), happen during(9), affect(8), be said in(8), be(7), be on(6), be found in(5), originate in(5), be experienced by(5), reside in(5), be held by(5), be done in(4), be made in(4), start in(4), be provided by(4), emerge from(3), begin at(3), be created by(3), be caused by(3), begin in(3), come in(3), be related to(3), be believed by(3), be made by(3), happen to(3), happen on(3), have(2), divide(2), include(2), be said(2), inhabit(2), characterize(2), support(2), happen(2), concern(2), be with(2), exist in(2), go into(2), be situated in(2), be within(2), be played during(2), dwell in(2), appear in(2), plague(2), belong to(2), reside on(2), form in(2), be sponsored by(2), emanate from(2), occur to(2), be consumed at(2), pertain to(2), be situated on(2), go at(2), be formed in(2), revolve around(1), influence(1), fly during(1), occur within(1), encompass(1), disrupt(1), be built in(1), result in(1), arise in(1), be during(1), occur(1), begin during(1), be by(1), be suffered by(1), exist on(1), crop up in(1), contain(1), be at(1), cover(1), result from(1), be recited in(1), require(1), effect(1), be defined by(1), afflict(1), be followed by(1), turn(1), befall(1), fly at(1), split(1), bother(1), serve(1), exist within(1), indicate(1), be of(1), come on(1), begin with(1), benefit(1), lie across(1), appear on(1), utilize(1), start at(1), be considered(1), be about(1), occur around(1), confuse(1), be measured at(1), fall in(1), be eaten around(1), happen by(1), be cultivated in(1), be had at(1), roam around(1), happpen during(1), come at(1), be offered by(1), encircle(1), be spoken in(1), be chanted in(1), be recited during(1), protect(1), be enjoyed in(1), mar(1), happen for(1), terminate in(1), purport(1), be housed in(1), comprise(1), enhance(1), be formed on(1), be believed in by(1), surprise(1), form within(1), be located on(1), shape(1), be led by(1), experience in(1), form at(1), eat in(1), be upheld by(1), be spawned in(1), hold across(1), sit in(1), . . .

Web: include(603), affect(319), live in(302), be(141), plague(128), bring(108), arise in(90), confront(79), lie beneath(69), be in(62), occur in(61), cause(61), run in(50), come from(49), have(47), involve(47), simmer beneath(44), leave(40), make(33), threaten(33), exist in(31), lurk beneath(30), face(29), extend into(29), move from(28), lead(28), arise within(28), leave at(28), run(28), beset(27), follow(27), keep(26), do(24), destroy(24), tear(23), put(23), force(23), impact(23), lay beneath(23), begin in(22), simmer under(22), be under(20), be faced by(20), represent(19), hold(19), devastate(18), be adopted by(18), be below(18), divide(17), work in(17), flee(17), simmer below(17), occur within(16), know(16), be from(16), arise from(16), be held(16), clean(16), break(15), bubble beneath(15), result in(14), concern(14), work(14), work for(14), exist for(14), be during(13), characterize(13), push(13), mean(13), populate(13), exist beneath(13), arise for(13), be read(13), come into(13), require(12), boil beneath(12), haunt(12), start(12), be solved by(12), touch(12), exist below(12), disrupt(11), be at(11), inhabit(11), be considered(11), dwell in(11), serve(11), live outside(11), build(11), drive(11), seethe beneath(11), be set by(11), ravage(11), make up(11), love(11), grow up in(11), allow(11), arrive at(11), give(11), be offered(11), exist within(10), be recited(10), burden(10), reflect(10), inhibit(10), run through(10), prevent(10), end(10), take(10), ensnare(10), exist at(10), provide(10), be beneath(10), create(10), be said(9), afflict(9), happen in(9), hurt(9), mean that(9), be called(9), run into(9), stretch into(9), mark(9), challenge(9), be generated by(9), visit(9), be located in(8), start in(8), bubble below(8), open in(8), remain under(8), lie below(8), prevail in(8), dog(8), volk for(8), suggest that(8), effect(7), be played in(7), occur at(7), use(7), arrive(7), lie under(7), be encountered in(7), know that(7), reduce(7), remain beneath(7), exist underneath(7), constitute(7), strain(7), care about(7), . . .

0.36 NOMINALIZATION:PRODUCT

Humans: be made by(41), contain(11), come from(11), be for(10), consist of(9), involve(9), ask for(7), include(7), request(7), be(6), be about(6), pay for(5), be made for(5), affect(5), concern(5), describe(5), be written by(4), be on(4), appear on(4), be made of(4), refer to(4), decorate(3), be of(3), bring in(3), fund(3), shorten(3), be made up of(3), comprise(2), supply(2), provide(2), create(2), distribute(2), define(2), be related to(2), be comprised of(2), be given to(2), cut(2), be done to(2), be created by(2), deal with(2), happen to(2), be made from(2), be performed by(2), help(1), cover(1), support(1), require(1), lower(1), serve(1), be used for(1), want(1), utilize(1), reduce(1), revolve around(1), be preformed on(1), be held by(1), go around(1), about(1), beautify(1), detail(1), emerge from(1), emanate(1), decrease(1), groom(1), be located in(1), help pay(1), neaten(1), be displayed on(1), be composed of(1), criticise(1), emaqnate from(1), be found in(1), be placed on(1), compose of(1), be generated of(1), come down from(1), be generated by(1), regard(1), demand for(1), discount(1), exist with(1), tell about(1), be handed down from(1), be passed by(1), cut from(1), be received by(1), trim(1), be made about(1), mitigate(1), pertain to(1), be funded by(1), praise(1), assist with(1), explain(1), be concerned with(1), originate from(1), stem from(1), be given by(1), come for(1), work on(1), subsidize(1), supplement(1), showcase(1), be created for(1), cut back(1), be published on(1), insult(1), relocate(1), evaluate(1), pay towards(1), be made on(1).

Web: include(99), allow(93), affect(65), be(58), appear on(42), be made by(36), make(36), grace(33), have(24), provide(19), be used for(17), consist of(16), become(16), be on(15), feature(13), help(11), require(10), involve(10), do(10), be captured on(9), design(9), limit(9), go against(9), utilize(8), reduce(8), find(8), keep(7), maintain(7), increase(7), be featured on(7), feature on(7), double as(7), be for(6), push(6), mean that(6), be derived from(6), use(6), be criticized by(6), comprise(5), create(5), be used as(5), be supported by(5), incorporate(5), embrace(5), rest with(5), constitute(5), restrict(5), adorn(5), flow around(5), distinguish(5), deploy(5), pay for(4), be made for(4), lower(4), be about(4), donate(4), eat(4), result in(4), handle(4), be used on(4), secure(4), find that(4), increase for(4), be sent between(4), minimize(4), enable(4), differ from(4), range from(4), lead(4), impact(4), be dispersed among(4), brace(4), be swiped from(4), engage(4), decorate(3), be of(3), define(3), encourage(3), bring(3), motivate(3), coordinate with(3), dismay(3), be dominated by(3), be incorporated into(3), complement(3), be purchased through(3), present(3), let(3), exclude(3), be located under(3), indulge in(3), be approved by(3), leave(3), influence(3), handwritten(3), embellish(3), publish(3), require that(3), be shared between(3), hold(3), replace(3), open(3), mean(3), preclude(3), place(3), be faced by(3), ask for(2), be written by(2), want(2), contain(2), come from(2), serve(2), package(2), jeopardize(2), render(2), ensure at(2), concur with(2), relate(2), approximate(2), stun(2), represent(2), be stamped on(2), intoxicate(2), flop(2), be mounted under(2), benefit(2), appear in(2), offset(2), be based on(2), borrow(2), be fed through(2), traverse(2), be ironed out at(2), permit(2), send(2), span(2), remove(2), divide(2), appear upon(2), bias(2), reflect(2), form(2), offer(2), run against(2), be depicted on(2), teach(2), hinge(2), double after(2), extend across(2), account for(2), ...

0.27 ABOUT

Humans: involve(81), be about(71), concern(33), be caused by(17), discuss(14), describe(11), deal with(10), be related to(9), relate to(9), pertain to(8), contain(7), focus on(7), feature(7), be over(7), refer to(6), have(5), come from(5), affect(5), talk about(4), explain(4), be due to(4), occur in(3), include(3), tell of(3), cover(3), support(3), speak of(3), be(3), be found in(3), be from(3), occur with(2), revolve around(2), be written about(2), center around(2), result from(2), tell about(2), require(2), be on(2), govern(2), cause(2), show(2), portray(2), define(2), be against(2), stem from(2), depend upon(1), center on(1), ban(1), regulate(1), detail(1), emerge from(1), sing about(1), surround(1), profess(1), write on(1), be created by(1), address(1), teach(1), regard(1), review(1), impact(1), be inspired by(1), approve(1), provide(1), tell(1), prohibit(1), study(1), speak about(1), justify(1), be concerned with(1), be of(1), use(1), follow(1), arise from(1), be for(1), be based on(1), complicate(1), express(1), be caused for(1), arise over(1), be made because of(1), dictate(1), overuse(1), be answered by(1), turn in(1), be relative to(1), belie to(1), reference(1), be fought over(1), be created as(1), terminate in(1), talk of(1), nourish(1), question(1), teach about(1), oppose(1), occur because of(1), be characterized by(1), be precipitated by(1), be pro(1), feed(1), be because of(1), highlight(1), be incited because of(1), imply(1), be regarding(1), be interested in(1), originate from(1), decide on(1), restrict(1), simulate(1), be taken regarding(1), be rooted in(1), detail plan for(1), concers(1), argue over(1), squeeze(1), argue about(1), be fought because of(1), happen because of(1), lie with(1), be made on(1).

Web: impose(476), be(273), make(171), affect(168), follow(147), involve(136), include(107), cover(103), raise(95), reduce(94), have(93), deal with(78), feature(75), be about(74), focus on(71), tell of(68), do(67), describe(64), pay(62), levy(61), celebrate(54), require(51), relate(51), recount(49), express(48), be covered by(44), combine(42), increase(42), tell(35), speak of(34), allow(34), chronicle(34), establish(32), go(32), continue(31), cut(31), place(31), explore(30), reflect(27), lower(27), go beyond(27), create(26), say(25), provide for(24), get(23), concern(22), cause(22), govern(21), tell about(21), put(21), revolve around(20), demand(20), talk about(19), mean(19), show(18), depict(18), present(17), reveal(17), regulate(16), teach(16), arise from(16), provide(16), bring(16), compare(16), bear upon(16), address(15), detail(15), prohibit(15), result in(15), give(15), make up(15), introduce(15), depend upon(14), eliminate(14), be collected in(14), promote(14), be known as(14), look at(14), offer(13), need(13), limit(13), contain(12), set(12), write(12), collect(12), represent(11), influence(11), call(11), be filled with(11), charge(11), conflict with(11), become(11), specialize in(11), result from(10), impact(10), define(10), be in(10), inspire(10), consider(10), examine(10), appear in(10), be influenced by(9), equalize(9), draw from(9), reflect on(9), be addressed in(9), embody(9), send(9), differ from(9), sing i(9), go down in(8), prevent(8), be within(8), go in(8), permit(8), narrate(8), capture(8), trace(8), seraphs(8), accompany(8), phase out(8), smell with(8), be told of(8), avoid(8), focus upon(8), epitomize(8), generate(8), lads(8), be based on(7), ban(7), regard(7), fall in(7), encompass(7), violate(7), push(7), distinguish between(7), convey(7), promise(7), be associated with(7), envisage(7), say that(7), repeal(7), originate in(7), supersede(7), enforce(7), authorize(7), emerge from(6), be of(6), explain(6), center on(6), . . .

0.23 NOMINALIZATION:ACT

Humans: be made by(56), involve(44), come from(19), occur in(12), affect(11), be done by(11), prevent(9), happen to(8), remove(6), happen in(6), include(5), concern(5), stimulate(5), go into(5), regulate(4), be performed by(4), be used by(4), eliminate(4), be caused by(4), explain(4), help(3), divide(3), be in(3), be for(3), occur with(3), emerge from(3), stop(3), reproduce(3), be of(3), limit(2), interpret(2), delete(2), resemble(2), target(2), be done on(2), examine(2), generate from(2), reclaim(2), consist of(2), be initiated by(2), study(2), be performed on(2), be applied to(2), restart(2), refer to(2), soothe(2), hurt(1), occur within(1), facilitate(1), get into(1), touch(1), restore(1), inhibit(1), be suffered by(1), form(1), be made for(1), control(1), assist(1), face(1), support(1), present(1), be directed at(1), be perpetrated by(1), provide(1), look at(1), replace(1), prohibit(1), save(1), be perpetrated on(1), split(1), occur through(1), change(1), use(1), focus on(1), produce(1), look like(1), feature(1), reduce(1), be about(1), investigate(1), be held by(1), revive(1), be conducted by(1), be provided for(1), create baby(1), be related to(1), divide up(1), exist for(1), accommodate(1), encompass(1), be meted out by(1), retrieve(1), be taken by(1), impede(1), free up(1), relax(1), be undertaken by(1), be promulgated by(1), signal(1), be composed of(1), cut though(1), be formed by(1), hatch(1), vibrate(1), be brought on by(1), be accomplished by(1), be made of(1), deindustrializes(1), palpate(1), be done to(1), get back(1), be created by(1), occur by(1), be started(1), berth(1), harbor(1), recover(1), leave out(1), accept(1), concern with(1), be designed for(1), occur to(1), be meant for(1), concentrate on(1), discuss(1), repress(1), be directed to(1), erase(1), go through(1), invade(1), be provided by(1), be from(1), receive(1), dissect(1), originate from(1), depict(1), multiply(1), be directed toward(1), be transmitted by(1), lead into(1), be forced by(1), describe(1), be specific to(1), take back(1), plague(1), map(1), massage(1), happen between(1), open to(1), bring about(1), . . .

Web: affect(711), produce(173), allow(104), give(95), result in(78), be(67), occur in(55), cause(54), involve(51), generate(51), require(47), support(45), place(45), bring(41), benefit(41), be made by(40), divide(38), put(38), make(38), leave(37), harm(32), enable(30), drive(29), have(27), turn(25), teach(24), include(22), impact(21), prevent(20), limit(20), remove(20), be inflicted upon(20), form(18), provide(18), consider(17), hurt(16), help(16), permit(16), occur within(15), be perpetrated on(14), expose(14), destroy(14), separate(13), deprive(13), assist(12), be inflicted on(12), be witnessed by(12), treat(12), abuse(12), commit(11), scar(11), reflect(11), be associated with(11), protect(11), promote(11), create(11), delete(10), prohibit(10), force(10), live with(10), acne(10), convert(10), ongoing in(10), be made for(9), encourage(9), exclude(9), take(9), lead(9), be heaped upon(9), render(9), occur after(9), find(9), be observed in(9), say(9), whip(9), be for(8), uphold(8), segregate(8), undergo(8), be perpetrated against(8), be confirmed by(8), be made about(8), admit(8), keep(8), improve(8), withdraw(8), shape(8), connect(8), occur at(8), be in(7), touch(7), come from(7), kill(7), refer(7), be heaped on(7), be held after(7), be found in(7), yield(7), be alleged in(7), damage(7), send(7), confront(7), invalidate(7), be upheld by(7), be correlated with(7), serve(7), indicate that(6), enhance(6), be presumed in(6), discourage(6), be followed by(6), influence(6), reflect on(6), be shaped by(6), parent(6), be reported in(6), be faced by(6), be expressed by(6), prepare(6), endanger(6), graduate(6), find that(6), be based on(6), contain(6), characterize(6), effect(6), injure(6), replace(5), concern(5), look at(5), mean that(5), befall(5), compel(5), be perceived by(5), do(5), own(5), get(5), be accompanied by(5), hinder(5), occur before(5), attract(5), be compared with(5), say that(5), talk about(5), nourish(5), be made in(5), be visited upon(5), introduce(5), authorize(5), upset(5), . . .

0.19 NOMINALIZATION:PATIENT

Humans: be made by(31), work for(18), come from(16), be created by(10), work at(9), be from(5), work in(4), be trained by(4), be produced by(3), be in(3), be hired by(3), live in(2), be trained in(2), be designed by(2), be done by(1), be enrolled in(1), be taught by(1), be dismissed by(1), be denied by(1), be found in(1), be rejected in(1), focus on(1), appear in(1), be discarded by(1), go to(1), be refused from(1), plague(1), produce by(1), be engineered by(1), be made in(1), happen in(1), help out at(1), be disposed of by(1).

Web: leave(98), be employed by(72), graduate from(63), attend(50), work for(32), work at(29), serve(29), have(23), be enrolled in(22), retire from(22), join(19), allow(18), violate(11), resign from(11), work in(10), be employed at(10), separate from(10), be(10), be made by(9), complete(9), inform(8), wish(8), notify(8), support(7), be employed with(7), disregard(6), perceive(6), enable(6), access(6), represent(6), be with(6), participate in(6), terminate from(6), travel for(6), come from(5), be produced by(5), affect(5), be in(4), enroll in(4), be required by(4), sue(4), be at(4), live outside(4), be selected with(4), maximize(4), be terminated by(4), separate under(4), be found by(4), be developed by(4), be covered under(4), be performed by(4), be created by(3), begin(3), take(3), finish(3), staff(3), require(3), be insured by(3), be relieved from(3), be designated by(3), be funded by(3), start(3), graduate(3), possess(3), be discussed by(3), do(3), work within(3), enter(3), be reemployed by(3), serve at(3), authorize(3), reduce(3), be done by(2), be from(2), be hired by(2), be utilized by(2), relocate within(2), lack(2), meet(2), supervise(2), get(2), reimburse(2), work with(2), be authorized by(2), agree that(2), move(2), assist(2), enrol at(2), be involved in(2), focus(2), aid(2), pursue(2), land(2), drive(2), aide(2), accredit(2), be laid off by(2), put through(2), take up(2), be near(2), design(2), show(2), be retained by(2), participate(2), transitioning from(2), use(2), be separated from(2), attain(2), be accredited at(2), stimulate(2), live in(1), be split across(1), challenge(1), oversee(1), help(1), be presented by(1), be nominated by(1), include(1), transfer from(1), operate(1), steal from(1), capture(1), present(1), pass(1), love(1), be on(1), seek(1), run(1), involve(1), care about(1), be educated at(1), serve with(1), save(1), fill(1), remain with(1), teach at(1), save for(1), be employed within(1), become at(1), study at(1), be employed in(1), enrol in(1), work(1), be crated by(1), please(1), be reimbursed by(1), benefit from(1), exemplify(1), be concerned that(1), leave for(1), . . .

0.15 MAKE$_2$

Humans: be made of(191), be made from(61), contain(31), be comprised of(22), be composed of(22), consist of(22), be made up of(17), be(9), have(6), house(6), look like(5), come from(5), be formed from(4), include(4), taste like(4), be made out of(4), be formed of(4), belong to(4), resemble(3), involve(3), use(3), utilize(3), be printed on(2), hold(2), be cast from(2), provide(2), be created from(2), incorporate(2), run on(2), run through(2), be made with(2), cut(2), be of(2), comprise(2), be in(2), be manufactured from(2), carry(2), be occupied by(2), be inhabited by(2), be on(1), be connected by(1), represent(1), be constructed of(1), seem like(1), bring(1), be constructed from(1), be supplied by(1), be fashioned from(1), be like(1), work with(1), train(1), be composed from(1), be from(1), descend from(1), be formed by(1), be shown on(1), work on(1), form(1), shape(1), connect(1), be carried on(1), be covered in(1), support(1), link(1), be filled with(1), be formulated of(1), appear like(1), broadcast use(1), be comprised by(1), emerge from(1), embody(1), build from(1), be headed by(1), emulate(1), be played on(1), store(1), be owned by(1), be fabricated from(1), originate from(1), make from(1), be relayed by(1), swim in(1), program(1), carve away(1), castle(1), be created of(1), be build by(1), be created by(1), be all(1), be turned in(1), be ruled by(1), be produced on(1), shelter(1), reach through(1), loop like(1).

Web: include(1037), be(552), consist of(261), use(161), have(146), allow(119), contain(88), be made of(83), be of(79), be comprised of(76), be composed of(72), meet with(67), select(64), provide(63), advise(62), assist(62), guide(60), bring(57), affect(56), work with(54), evaluate(54), represent(53), need(53), use in(52), be cast in(52), require(51), review(46), involve(36), supervise(35), comprise(34), serve(34), examine(31), encourage(29), cause(28), make(27), oversee(26), give(26), inform(25), be made from(24), approve(23), wear(23), place(21), help(21), connect(20), recommend(20), look like(19), run(19), seek(19), become(19), run over(18), look through(18), interview(17), nominate(17), be hidden behind(17), employ(17), taste like(16), hear(16), remove(16), compete with(16), match(16), recognize(16), admit(16), be chosen by(16), be corrected with(16), exist on(15), consider(15), be in(14), combine(14), look for(14), choose(14), support(13), cut(13), fall into(13), be used in(13), break(13), be chaired by(13), compete in(13), educate(13), promote(12), move(12), benefit(12), rely on(12), operate(11), be fitted with(11), lead(11), be spent on(11), address(11), accept(11), lay(11), consult with(11), be broadcast via(11), be based on(11), incorporate(10), be made with(10), exclude(10), replace(10), welcome(10), put(10), assign(10), determine whether(10), be corrected by(10), be replaced with(10), bring together(10), prepare(9), serve as(9), identify(9), act as(9), operate over(9), be covered with(9), get(9), decide whether(9), ask(9), graduate(9), train(8), run on(8), be covered in(8), bring back(8), be called(8), offer(8), test(8), be dusted with(8), be selected by(8), gleam like(8), suck in(8), be among(8), see(8), transport(8), compete against(8), compose of(8), build with(8), assess(8), deal with(8), be sold in(8), say(8), be filled with(7), hold(7), read(7), dismiss(7), keep(7), be represented by(7), be wrapped in(7), question(7), catapult(7), injure(7), feature(7), be replaced by(7), direct(7), eliminate(7), recommend whether(7), be acquired by(7), ...

0.11 HAVE$_2$

Humans: come from(48), belong to(48), be owned by(27), surround(25), be possessed by(11), be on(10), cover(9), be in(7), involve(7), be of(6), be found on(6), protect(5), be held by(4), be had by(4), grow on(3), measure(3), be exercised by(3), be found in(3), affect(3), concern(3), hold(3), be within(3), encompass(2), reach(2), be located in(2), enclose(2), attend(2), include(2), form(2), be generated by(2), fit(2), be from(2), lie within(2), go to(2), relate to(2), be exerted by(2), come off(2), be inside of(2), occur with(1), join(1), weigh(1), limit(1), have(1), occur within(1), be set for(1), carry(1), motivate(1), emerge from(1), be controlled by(1), frustrate(1), be kept by(1), help(1), be inside(1), make up(1), encircle(1), be used by(1), contain(1), characterize(1), circle(1), be kept in(1), result from(1), survive(1), accept(1), worry(1), drive(1), build up(1), fall from(1), be made by(1), derive from(1), participate in(1), inspire(1), be caused by(1), burden(1), originate from(1), remain with(1), reside with(1), shed from(1), trouble(1), reside in(1), be(1), be encountered by(1), be seized by(1), plague(1), evaluate(1), determine(1), define(1), be about(1), be reached by(1), be involved with(1), be expressed by(1), devolve from(1), be contained within(1), be linked to(1), be related to(1), be handed down in(1), of(1), emanate(1), lock in(1), reach to(1), be layered on(1), reproduce(1), be given to(1), be wielded by(1), occur(1), be surrounded by(1), be made of(1), be compared to(1), be measured for(1), be shown by(1), center(1), belong with(1), be worn by(1), arise out of(1), face challenge(1), be due to(1), be attributed to(1), be yielded by(1), be ascribed to(1), be passed down within(1), happen to(1), be like(1), interact at(1), border around(1), be made from(1), be included with(1), be measured on(1), pass through(1), surronds(1).

Web: be(3805), surround(709), require(598), affect(529), enable(388), allow(385), be owned by(287), leave(237), represent(230), protect(218), confront(217), prevent(204), encircle(203), divide(185), join(172), cause(170), lead(161), make(139), ask(120), be in(117), have(113), be acquired by(111), help(109), keep(106), encourage(103), interfere with(99), bring(96), engage(93), force(86), arise between(82), be held by(80), become(79), be used by(76), support(76), hold(75), attend(74), enclose(72), contract(72), challenge(72), face(72), attract(67), serve(62), include(58), be adopted by(58), separate(58), take(55), provide(54), involve(52), impact(52), motivate(51), want(50), switch(48), oppress(47), give(47), turn(47), drive(45), draw(44), be purchased by(41), ring(41), assist(41), plague(40), increase(40), deprive(40), be solved by(40), arise from(39), change(39), be taken by(39), come from(38), kill(37), form(36), prepare(36), be granted by(36), fit(35), put(35), defend(35), meet(35), quit(35), arise for(35), carry(34), be of(34), arise with(34), expose(33), test(33), impede(32), be expelled from(32), benefit(32), arise in(31), hinder(30), transform(29), be selected by(28), resign from(28), be affiliated with(28), stand between(27), book(27), threaten(27), be used with(27), encompass(26), move(26), be leased from(26), mean that(26), destroy(26), oppose(25), embarrass(25), be sold by(25), be vested in(24), be acquired from(23), be built around(23), set(23), influence(22), perplex(22), be that(22), remain in(22), appear on(22), exploit(21), arise during(21), defect from(21), reflect(21), be confiscated by(21), control(21), permit(21), be with(21), prompt(21), be amassed by(21), exist in(21), guard(20), save(20), be registered with(20), disagree with(20), strengthen(20), prohibit(20), result from(19), define(19), be found on(19), limit(19), hurt(19), run through(19), be included in(19), be given by(19), impair(19), connect(19), be delegated by(19), be set by(18), introduce(18), sway(18), push(18), be appointed by(18), need(18), enhance(18), be made(18), be embedded in(18), . . .