A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books

Yoav Goldberg
Bar Ilan University*
yoav.goldberg@gmail.com

Jon Orwant
Google Inc.
orwant@google.com

Abstract

We created a dataset of syntactic-ngrams (counted dependency-tree fragments) based on a corpus of 3.5 million English books. The dataset includes over 10 billion distinct items covering a wide range of syntactic configurations. It also includes temporal information, facilitating new kinds of research into lexical semantics over time. This paper describes the dataset, the syntactic representation, and the kinds of information provided.

1 Introduction

The distributional hypothesis of Harris (1954) states that properties of words can be captured based on their contexts. The consequences of this hypothesis have been leveraged to a great effect by the NLP community, resulting in algorithms for inferring syntactic as well as semantic properties of words (see e.g. (Turney and Pantel, 2010; Baroni and Lenci, 2010) and the references therein).

In this paper, we describe a very large dataset of *syntactic*-ngrams, that is, structures in which the contexts of words are based on their respective position in a syntactic parse tree, and not on their sequential order in the sentence: the different words in the ngram may be far apart from each other in the sentence, yet close to each other syntactically. See Figure 1 for an example of a syntactic-ngram.

The utility of syntactic contexts of words for constructing vector-space models of word meanings is well established (Lin, 1998; Lin and Pantel, 2001; Padó and Lapata, 2007; Baroni and Lenci, 2010). Syntactic relations are successfully used for modeling selectional preferences (Erk and Padó, 2008; Erk et al., 2010; Ritter et al., 2010; Séaghdha, 2010), and dependency paths are also used to infer binary relations between words (Lin and Pantel, 2001; Wu and Weld, 2010). The use of syntactic-ngrams holds promise also for improving the accuracy of core NLP tasks such as syntactic language-modeling (Shen et al., 2008) and syntactic-parsing (Chen et al., 2009; Sagae and Gordon, 2009; Cohen et al., 2012), though most successful attempts to improve syntactic parsing by using counts from large corpora are based on sequential rather than syntactic information (Koo et al., 2008; Bansal and Klein, 2011; Pitler, 2012), we believe this is because large-scale datasets of syntactic counts are not readily available. Unfortunately, most work utilizing counts from large textual corpora does not use a standardized corpora for constructing their models, making it very hard to reproduce results and challenging to compare results across different studies.

Our aim in this work is not to present new methods or results, but rather to provide a new kind of a large-scale (based on corpora about 100 times larger than previous efforts) high-quality and standard resource for researchers to build upon. Instead of focusing on a specific task, we aim to provide a flexible resource that could be adapted to many possible tasks.

Specifically, the contribution of this work is in creating a dataset of syntactic-ngrams which is:

- Derived from a very large (345 billion words) corpus spanning a long time period.
- Covers a wide range of syntactic phenomena and is adaptable to many use cases.
- Based on state-of-the-art syntactic processing in a modern syntactic representation.
- Broken down by year of occurrence, as well as some coarse-grained regional and genre distinctions (British, American, Fiction).
- Freely available for non-commercial use. [1]

*Work performed while at Google.

[1]The dataset is made publicly available under the Creative Commons Attribution-Non Commercial ShareAlike 3.0 Unported License: http://creativecommons.org/licenses/by-nc-

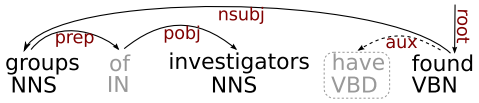

Figure 1: A syntactic ngram appearing 112 times in the *extended-biarcs* set, which include structures containing three content words (see Section 4). Grayed items are non-content words and are not included in the word count. The dashed auxiliary "have" is a functional marker (see Section 3), appearing only in the *extended-** sets.

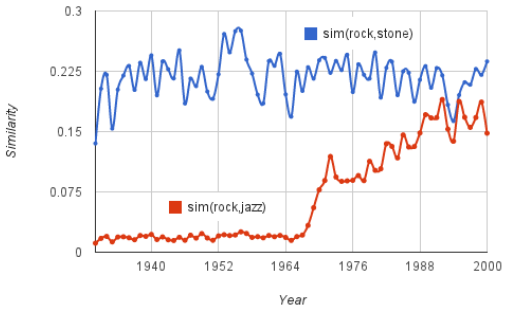

Figure 2: Word-similarity over time: The word "rock" starts to become similar to "jazz" around 1968. The plot shows the cosine similarity between the immediate syntactic contexts of of the word "rock" in each year, to the immediate syntactic contexts of the words "jazz" (in red) and "stone" (in blue) aggregated over all years.

After describing the underlying syntactic representation, we will present our definition of a syntactic-ngram, and detail the kinds of syntactic-ngrams we chose to include in the dataset. Then, we present details of the corpus and the syntactic processing we performed.

With respect to previous efforts, the dataset has the following distinguishing characteristics:

Temporal Dimension A unique aspect of our dataset is the temporal dimension, allowing inspection of how the contexts of different words vary over time. For example, one could examine how the meaning of a word evolves over time by looking at the contexts it appears in within different time periods. Figure 2 shows the cosine similarity between the word "rock" and the words "stone" and "jazz" from year 1930 to 2000, showing that rock acquired a new meaning around 1968.

Large syntactic contexts Previous efforts of providing syntactic counts from large scale corpora (Baroni and Lenci, 2010) focus on relations between two content words. Our dataset include structures covering much larger tree fragments, some of them including 5 or more content words. By including such structures we hope to encourage research exploring higher orders of interactions, for example modeling the relation between adjectives of two conjoined nouns, the interactions between subjects and objects of verbs, or fine-grained selectional preferences of verbs and nouns.

A closely related effort to add syntactic annotation to the books corpus is described in Lin et al. (2012). That effort emphasize an interactive query interface covering several languages, in which the underlying syntactic representations are linear-ngrams enriched with universal part-of-speech tags, as well as first order unlabeled dependencies. In contrast, our emphasis is not on an easy-to-use query interface but instead a useful and flexible resource for computational-minded researchers. We focus on English and use finer-grained English-specific POS-tags. The syntactic analysis is done using a more accurate parser, and we provide counts over *labeled* tree fragments, covering a diverse set of tree-fragments many of which include more than two content words.

Counted Fragments instead of complete trees While some efforts provide complete parse trees from large corpora (Charniak, 2000; Baroni et al., 2009; Napoles et al., 2012), we instead provide counted tree fragments. We believe that our form of aggregate information is of more immediate use than the raw parse trees. While access to the parse trees may allow for somewhat greater flexibility in the kinds of questions one could ask, it also comes with a very hefty price tag in terms of the required computational resources: while counting seems trivial, it is, in fact, quite demanding computationally when done on such a scale, and requires a massive infrastructure. By lifting this burden of NLP researchers, we hope to free them to tackle interesting research questions.

2 Underlying Syntactic Representation

We assume the part-of-speech tagset of the Penn Treebank (Marcus et al., 1993). The syntactic representation we work with is based on dependency-grammar. Specifically, we use labeled dependency trees following the "basic" variant of the Stanford-dependencies scheme (de Marneffe and Manning, 2008b; de Marneffe and Manning, 2008a).

Dependency grammar is a natural choice, as it emphasizes individual words and explicitly models the connections between them. Stanford dependencies are appealing because they model relations between content words directly, without intervening functional markers (so in a construction

sa/3.0/legalcode.

such as "wanted to know" there is a direct relation (wanted, know) instead of two relation (wanted, to) and (to, know). This facilitates focusing on meaning-bearing content words and including the maximal amount of information in an ngram.

3 Syntactic-ngrams

We define a syntactic-ngram to be a rooted connected dependency tree over k words, which is a subtree of a dependency tree over an entire sentence. For each of the k words in the ngram, we provide information about the word-form, its part-of-speech, and its dependency relation to its head. The ngram also indicates the relative ordering between the different words (the order of the words in the syntactic-ngram is the same as the order in which the words appear in the underlying sentence) but not the distance between them, nor an indication whether there is a missing material between the nodes. Examples of syntactic-ngrams are provided in Figures 1 and 3.

Content-words and Functional-markers We distinguish between *content-words* which are meaning bearing elements and *functional-markers*, which serve to add polarity, modality or definiteness information to the meaning bearing elements, but do not carry semantic meaning of their own, such as the auxiliary verb "have" in Figure 1. Specifically, we treat words with a dependency-label of det, poss, neg, aux, auxpass, ps, mark, complm and prt as functional-markers. With the exception of poss, these are all closed-class categories. All other words except for prepositions and conjunctions are treated as content-words. A syntactic-ngram of order n includes exactly n content words. It may optionally include all of the functional-markers that modify the content-words.

Conjunctions and Prepositions Conjunctions and Prepositions receive a special treatment. When a coordinating word ("and", "or", "but") appears as part of a conjunctive structure (e.g. "X, Y, and Z"), it is treated as a non-content word. Instead, it is always included in the syntactic-ngrams that include the conjunctive relation it is a part of, allowing to differentiate between the various kinds of conjunctions. An example is seen in Figure 3d, in which the relation conj(efficient, effective) is enriched with the coordinating word "or". When a coordinating word does not explicitly take part in a conjunction relation (e.g. "But, ...") it is treated as a content word.

When a preposition is part of a prepositional modification (i.e. in the middle of the pair (prep, pcomp) or (prep, pobj)), such as the word "of" in Figures 1 and 3h and the word "as" in Figure 3e, it is treated as a non-content word, and is always included in a syntactic-ngram whenever the words it connects are included. In cases of ellipsis or other cases where there is no overt pobj or pcomp ("he is hard to deal with") the preposition is treated as a content word.[2]

Multiword Expressions Some multiword expressions are recognized by the parser. Whenever a content word in an ngram has modifiers with the mwe relation, they are included in the ngram.

4 The Provided Ngram Types

We aimed to include a diverse set of relations, with maximal emphasis on relations between content-bearing words, while still retaining access to definiteness, modality and polarity if they are desired. The dataset includes the following types of syntactic structures:

nodes (47M items) consist of a single content word, and capture the syntactic role of that word (as in Figure 3a). For example, we can learn that the pronoun "he" is predominantly used as a subject, and that "help" as a noun is over 4 times more likely to appear in object than in subject position.

arcs (919M items) consist of two content words, and capture direct dependency relations such as "subject of", "adverbial modifier of" and so on (see Figure 3c,3d for examples). These correspond to "dependency triplets" as used in Lin (1998) and most other work on syntax-based semantic similarity.

biarcs (1.78B items) consist of three content words (either a word and its two daughters, or a child-parent-grandparent chain) and capture relations such as "subject verb object", "a noun and

[2] This treatment of prepositions and conjunction is similar to the "collapsed" variant of Stanford Dependencies (de Marneffe and Manning, 2008a), in which preposition- and conjunction-words do not appear as nodes in the tree but are instead annotated on the dependency label between the content words they connect, e.g. prep_with(saw, telescope). However, we chose to represent the preposition or conjunction as a node in the tree rather than moving it to the dependency label as it retains the information about the location of the function word with respect to the other words in the structure, is consistent with cases in which one of the content words is not present, and does not blow up the label-set size.

two adjectivial modifiers", "verb, object and adjectivial modifier of the object" and many others.

triarcs (1.87B items) consist of four content words (example in Figure 3f). The locality of the dependency representation causes this set of three-arcs structures to be large, sparse and noisy – many of the relations may appear random because some arcs are in many cases almost independent given the others. However, some of the relations are known to be of interest, and we hope more of them will prove to be of interest in the future. Some of the interesting relations include:
- modifiers of the head noun of the subject or object in an SVO construction: ((small,boy), ate, cookies), (boy, ate, (tasty, cookies)), and with abstraction: adjectives that a boy likes to eat: (boy, ate, (tasty, *))
- arguments of an embeded verb (said, (boy, ate, cookie)), (said, ((small, boy), ate))
- modifiers of conjoined elements ((small, boy) (young, girl)) , ((small, *) (young, *))
- relative clause constructions (boy, (girl, with-cookies, saw))

quadarcs (187M items) consist of 5 content words (example in Figure 3h). In contrast to the previous datasets, this set includes only a subset of the possible relations involving 5 content words. We chose to focus on relations which are attested in the literature (Padó and Lapata 2007; Appendix A), namely structures consisting of two chains of length 2 with a single head, e.g. ((small, boy), ate, (tasty, cookie)).

extended-nodes, extended-arcs, extended-biarcs, extended-triarcs, extended-quadarcs (80M, 1.08B, 1.62B, 1.71B, and 180M items) Like the above, but the functional markers of each content words are included as well (see examples in Figures 3b, 3e, 3g). These structures retain information regarding aspects such as modality, polarity and definiteness, distinguishing, e.g. "his red car" from "her red car", "will go" from "should go" and "a best time" from "the best time".

verbargs (130M items) This set of ngrams consist of verbs with all their immediate arguments, and can be used to study interactions between modifiers of a verb, as well as subcategorization frames. These structures are also useful for syntactic language modeling, as all the daughters of a verb are guaranteed to be present.

nounargs (275M items) This set of ngrams consist of nouns with all their immediate arguments.

Corpus	# Books	# Pages	# Sentences	# Tokens
All	3.5M	925.7M	17.6B	345.1B
1M	1M	291.1M	5.1B	101.3B
Fiction	817K	231.3M	4.7B	86.1B
American	1.4M	387.6M	7.9B	146.2B
British	423K	124.9M	2.4B	46.1B

Table 1: Corpora sizes.

verbargs-unlex, nounargs-unlex (114M, 195M items) Like the above, but only the head word and the top-1000 occurring words in the English-1M subcorpus are lexicalized – other words are replaced with a ⋆W⋆ symbol. By abstracting away from non-frequent words, we include many of the larger syntactic configurations that will otherwise be pruned away by our frequency threshold. These could be useful for inspecting fine-grained syntactic subcategorization frames.

5 Corpora and Syntactic Processing

The dataset is based on the English Google Books corpus. This is the same corpus used to derive the Google Books Ngrams, and is described in detail in Michel et al. (2011). The corpus consists of the text of 3,473,595 English books which were published between 1520 and 2008, with the majority of the content published after 1800. We provide counts based on the entire corpus, as well as on several subsets of it:
English 1M Uniformly sampled 1 million books.
Fiction Works of Fiction.
American English Books published in the US.
British English Books published in Britain.
The sizes of the different corpora are detailed in Table 1.

Counts Each syntactic ngram in each of the subcorpora is coupled with a corpus-level count as well as counts from each individual year. To keep the data manageable, we employ a frequency threshold of 10 on the corpus-level count.

Data Processing We ignored pages with over 600 white-spaces (which are indicative of OCR errors or non-textual content), as well as sentences of over 60 tokens. Table 1 details the sizes of the various corpora.

After OCR, sentence splitting and tokenization, the corpus went through several stages of syntactic processing: part-of-speech tagging, syntactic parsing, and syntactic-ngrams extraction.

Part-of-speech tagging was performed using a first order CRF tagger, which was trained on a

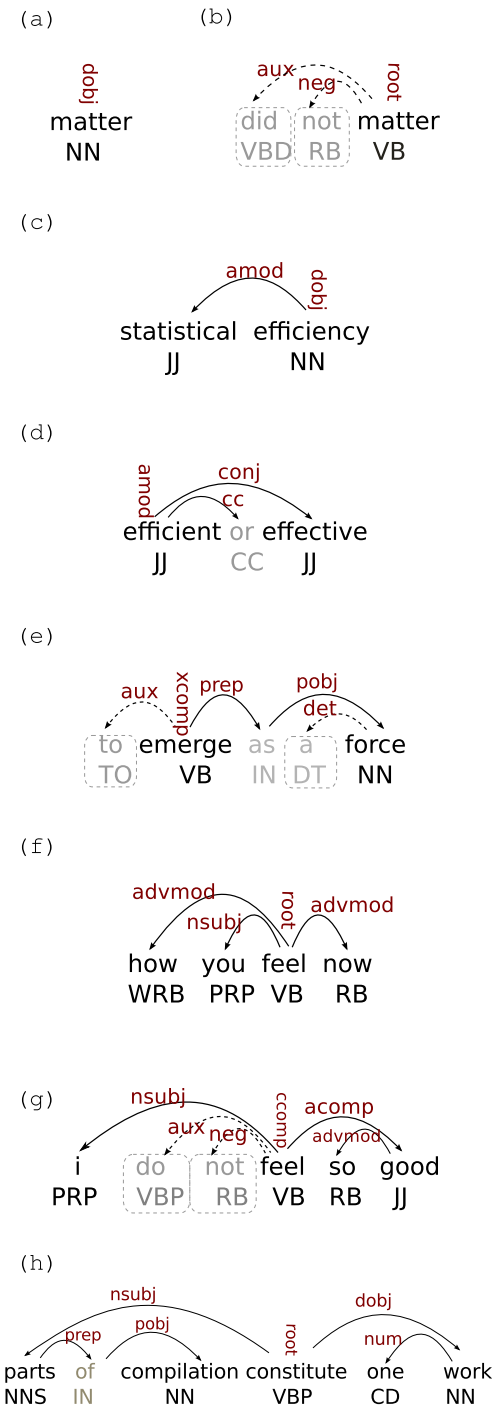

Figure 3: Syntactic-ngram examples. Non-content words are grayed, functional markers appearing only in the extended-* collections are dashed. (a) node (b) extended-node (c) arcs (d) arcs, including the coordinating word (e) extended-arcs, including a preposition (f) triarcs (g) extended-triarcs (h) quadarcs, including a preposition.

union of the Penn WSJ Corpus (Marcus et al., 1993), the Brown corpus (Kucera and Francis, 1967) and the Questions Treebank (Judge et al., 2006). In addition to the diverse training material, the tagger makes use of features based on word-clusters derived from trigrams of the Books corpus. These cluster-features make the tagger more robust on the books domain. For further details regarding the tagger, see Lin et al. (2012).

Syntactic parsing was performed using a re-implementation of a beam-search shift-reduce dependency parser (Zhang and Clark, 2008) with a beam of size 8 and the feature-set described in Zhang and Nivre (2011). The parser was trained on the same training data as the tagger after 4-way jack-knifing so that the parser is trained on data with predicted part-of-speech tags. The parser provides state-of-the-art syntactic annotations for English.[3]

6 Conclusion

We created a dataset of syntactic-ngrams based on a very large literary corpus. The dataset contains over 10 billion unique items covering a wide range of syntactic structures, and includes a temporal dimension.

The dataset is available for download at `http://storage.googleapis.com/books/syntactic-ngrams/index.html`

Acknowledgments

We would like to thank the members of Google's extended syntactic-parsing team (Ryan McDonald, Keith Hall, Slav Petrov, Dipanjan Das, Hao Zhang, Kuzman Ganchev, Terry Koo, Michael Ringgaard and, at the time, Joakim Nivre) for many discussions, support, and of course the creation and maintenance of an extremely robust parsing infrastructure. We further thank Fernando Pereira for supporting the project, and Andrea Held and Supreet Chinnan for their hard work in making this possible. Sebastian Padó, Marco Baroni, Alessandro Lenci, Jonathan Berant and Dan Klein provided valuable input that helped shape the final form of this resource.

[3]Evaluating the quality of syntactic annotation on such a varied dataset is a challenging task on its own right – the underlying corpus includes many different genres spanning different time periods, as well as varying levels of digitization and OCR quality. It is extremely difficult to choose a representative sample to manually annotate and evaluate on, and we believe no single number will do justice to describing the annotation quality across the entire dataset. On top of that, we then aggregate fragments and filter based on counts, further changing the data distribution. We feel that it is better not to provide any numbers than to provide inaccurate, misleading or uninformative numbers. We therefore chose not to provide a numeric estimation of syntactic-annotation quality, but note that we used a state-of-the-art parser, and believe most of its output to be correct, although we do expect a fair share of annotation errors as well.

References

Mohit Bansal and Dan Klein. 2011. Web-scale features for full-scale parsing. In *ACL*, pages 693–702.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Eugene Charniak. 2000. Bllip 1987-89 wsj corpus release 1. In *Linguistic Data Consortium*, Philadelphia.

Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Improving dependency parsing with subtrees from auto-parsed data. In *EMNLP*, pages 570–579.

Raphael Cohen, Yoav Goldberg, and Michael Elhadad. 2012. Domain adaptation of a dependency parser with a class-class selectional preference model. In *Proceedings of ACL 2012 Student Research Workshop*, pages 43–48, Jeju Island, Korea, July. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008a. Stanford dependencies manual. Technical report, Stanford University.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008b. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, Honolulu, HI. To appear.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proc. of ACL*, pages 497–504. Association for Computational Linguistics.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proc. of ACL*, pages 595–603.

Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.

Dekang Lin and Patrick Pantel. 2001. Dirt: discovery of inference rules from text. In *KDD*, pages 323–328.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *ACL (System Demonstrations)*, pages 169–174.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *AKBC-WEKEX Workshop at NAACL 2012*, June.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Emily Pitler. 2012. Attacking parsing bottlenecks with unlabeled data and relevant factorizations. In *ACL*, pages 768–776.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *ACL*, pages 424–434.

Kenji Sagae and Andrew S. Gordon. 2009. Clustering words by syntactic similarity improves dependency parsing of predicate-argument structures. In *IWPT*, pages 192–201.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *ACL*, pages 435–444.

Libin Shen, Jinxi Xu, and Ralph M. Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *ACL*, pages 577–585.

P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In *ACL*, pages 118–127.

Yue Zhang and Stephen Clark. 2008. A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. In *Proc. of EMNLP*, pages 562–571.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193.