Syntactic Annotations for the Google Books Ngram Corpus

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden,
Jon Orwant, Will Brockman and Slav Petrov*
Google Inc.
{yurilin,jbmichel,drerez,orwant,brockman,slav}@google.com

Abstract

We present a new edition of the Google Books Ngram Corpus, which describes how often words and phrases were used over a period of five centuries, in eight languages; it reflects 6% of all books ever published. This new edition introduces syntactic annotations: words are tagged with their part-of-speech, and head-modifier relationships are recorded. The annotations are produced automatically with statistical models that are specifically adapted to historical text. The corpus will facilitate the study of linguistic trends, especially those related to the evolution of syntax.

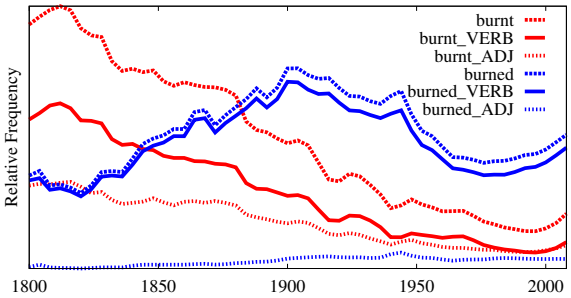


Figure 1: Usage frequencies of *burned* and *burnt* over time, showing that *burned* became the dominant spelling around 1880. Our new syntactic annotations enable a more refined analysis, suggesting that the crossing-point for the verb usage (*burned_VERB* vs. *burnt_VERB*) was decades earlier.

1 Introduction

The Google Books Ngram Corpus (Michel et al., 2011) has enabled the quantitative analysis of linguistic and cultural trends as reflected in millions of books written over the past five centuries. The corpus consists of words and phrases (i.e., ngrams) and their usage frequency over time. The data is available for download, and can also be viewed through the interactive Google Books Ngram Viewer at http://books.google.com/ngrams.

The sheer quantity of and broad historical scope of the data has enabled a wide range of analyses (Michel et al., 2011; Ravallion, 2011). Of course, examining raw ngram frequencies is of limited utility when studying many aspects of linguistic change, particularly the ones related to syntax. For instance, most English verbs are regular (their past tense is formed by adding -ed), and the few exceptions, known as irregular verbs, tend to regularize over the centuries (Lieberman et al., 2007). Figure 1 illustrates how *burned* gradually overtook *burnt*, becoming more frequent around 1880. Unfortunately, as a study of verb regularization, this analysis is skewed by a significant confound: both words can serve as either verbs (e.g., *the house burnt*) or adjectives (e.g., *the burnt toast*). Because many words have multiple syntactic interpretations, such confounds often limit the utility of raw ngram frequency data.

In this work we provide a new edition of the Google Books Ngram Corpus that contains over 8 million books, or 6% of all books ever published (cf. Section 3). Moreover, we include syntactic analysis in order to facilitate a fine-grained analysis of the evolution of syntax. Ngrams are annotated with part-of-speech tags (e.g., in the phrase *he burnt the toast*, *burnt* is a verb; in *the burnt toast*, *burnt* is an adjective) and head-modifier dependencies (e.g., in the phrase *the little black book*, *little* modifies *book*).

The annotated ngrams are far more useful for ex-

* Corresponding author.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 169–174,
Jeju, Republic of Korea, 8-14 July 2012.

amining the evolution of grammar and syntax. For our study of the regularization of the verb *burn*, the availability of syntactic annotations resolves the verb vs. adjective ambiguity in the original data, allowing us to only examine instances where *burnt* and *burned* appear as verbs. This more refined analysis suggests a crossover date for the frequency of the verb forms that is several decades earlier than the overall (verbs and adjectives) crossover.

We use state-of-the-art statistical part-of-speech taggers and dependency parsers to produce syntactic annotations for eight languages in the Google Books collection. The annotations consist of 12 language universal part-of-speech tags and unlabeled head-modifier dependencies. Section 4 describes the models that we used and the format of the annotations in detail. We assess the expected annotation accuracies experimentally and discuss how we adapt the taggers and parsers to historical text in Section 5. The annotated ngrams are available as a new edition of the Google Books Ngram Corpus; we provide some examples from the new corpus in Figure 3.

2 Related Work

Michel et al. (2011) described the construction of the first edition of the Google Books Ngram Corpus and used it to quantitatively analyze a variety of topics ranging from language growth to public health. The related Ngram Viewer has become a popular tool for examining language trends by experts and non-experts alike.

In addition to studying frequency patterns in the data, researchers have also attempted to analyze the grammatical function of the ngrams (Davies, 2011). Such endeavors are hampered by the fact that the Ngram Corpus provides only aggregate statistics in the form of ngram counts and not the full sentences. Furthermore, only ngrams that pass certain occurrence thresholds are publicly available, making any further aggregation attempt futile: in heavy tail distributions like the ones common in natural languages, the counts of rare events (that do not pass the frequency threshold) can have a large cumulative mass.

In contrast, because we have access to the full text, we can annotate ngrams to reflect the particular grammatical functions they take in the sentences they were extracted from, and can also account for the contribution of rare ngrams to otherwise frequent grammatical functions.

Language	#Volumes	#Tokens
English	4,541,627	468,491,999,592
Spanish	854,649	83,967,471,303
French	792,118	102,174,681,393
German	657,991	64,784,628,286
Russian	591,310	67,137,666,353
Italian	305,763	40,288,810,817
Chinese	302,652	26,859,461,025
Hebrew	70,636	8,172,543,728

Table 1: Number of volumes and tokens for each language in our corpus. The total collection contains more than 6% of all books ever published.

3 Ngram Corpus

The Google Books Ngram Corpus has been available at `http://books.google.com/ngrams` since 2010. This work presents new corpora that have been extracted from an even larger book collection, adds a new language (Italian), and introduces syntactically annotated ngrams. The new corpora are available in addition to the already existing ones.

3.1 Books Data

The new edition of the Ngram Corpus supports the eight languages shown in Table 1. The book volumes were selected from the larger collection of all books digitized at Google following exactly the procedure described in Michel et al. (2011). The new edition contains data from 8,116,746 books, or over 6% of all books ever published. The English corpus alone comprises close to half a trillion words. This collection of books is much larger than any other digitized collection; its generation required a substantial effort involving obtaining and manually scanning millions of books.

3.2 Raw Ngrams

We extract ngrams in a similar way to the first edition of the corpus (Michel et al., 2011), but with some notable differences. Previously, tokenization was done on whitespace characters and all ngrams occurring on a given page were extracted, including ones that span sentence boundaries, but omitting

Tag	English	Spanish	French	German	Russian[1]	Italian	Chinese	Hebrew
ADJ	other, such	mayor, gran	tous, même	anderen, ersten	все,этой	stesso, grande	大, 新	גדול, אחר
ADP	of, in	de, en	de, à	in, von	в, на	di, in	在, 对	ב, ל
ADV	not, when	no, más	ne, plus	auch, so	так, более	non, piú	不, 也	לא, כל
CONJ	and, or	y, que	et, que	und, daß	и, что	che, ed	和, 与	כי, ו
DET	the, a	la, el	la, les	der, die	-	la, il	这, 各	ה
NOUN	time, people	parte, años	temps, partie	Zeit, Jahre	его, он	parte, tempo	年, 人	ישראל, בית
PRON	it, I	que, se	qui, il	sich, die	-	che, si	他, 我	זה, הוא
VERB	is, was	es, ha	est, sont	ist, werden	было, был	é, sono	是, 有	היה, אין

Table 2: The two most common words for some POS tags in the new Google Books NGram Corpus for all languages.

ngrams that span page boundaries.

Instead, we perform tokenization and sentence boundary detection by applying a set of manually devised rules (except for Chinese, where a statistical system is used for segmentation). We capture sentences that span across page boundaries, and then extract ngrams only within sentences. As is typically done in language model estimation, we add sentence beginning (*_START_*) and end tokens (*_END_*) that are included in the ngram extraction. This allows us to distinguish ngrams that appear in sentence-medial positions from ngrams that occur at sentence boundaries (e.g., *_START_ John*).

3.3 Differences to the First Edition

The differences between this edition and the first edition of the Ngram Corpus are as follows: (i) the underlying book collection has grown substantially in the meantime; (ii) OCR technology and metadata extraction have improved, resulting in higher quality digitalization; (iii) ngrams spanning sentence boundaries are omitted, and ngrams spanning page boundaries are included. As a result, this new edition is not a superset of the first edition.

4 Syntactic Annotations

In addition to extracting raw ngrams, we part-of-speech tag and parse the entire corpus and extract syntactically annotated ngrams (see Figure 2). We use manually annotated treebanks of modern text (often newswire) as training data for the POS tagger and parser models. We discuss our approach to adapting the models to historical text in Section 5.

4.1 Part-of-Speech Tagging

Part-of-speech tagging is one of the most fundamental disambiguation steps in any natural language processing system. Over the years, POS tagging accuracies have steadily improved, appearing to plateau at an accuracy level that approaches human inter-annotator agreement (Manning, 2011). As we demonstrate in the next section, these numbers are misleading since they are computed on test data that is very close to the training domain. We therefore need to specifically adapt our models to handle noisy and historical text.

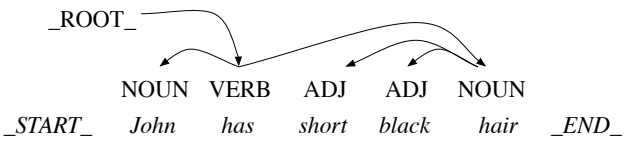
We perform POS tagging with a state-of-the-art[2] Conditional Random Field (CRF) based tagger (Lafferty et al., 2001) trained on manually annotated treebank data. We use the following fairly standard features in our tagger: current word, suffixes and prefixes of length 1, 2 and 3; additionally we use word cluster features (Uszkoreit and Brants, 2008) for the current word, and transition features of the cluster of the current and previous word.

To provide a language-independent interface, we use the universal POS tagset described in detail in Petrov et al. (2012). This universal POS tagset defines the following twelve POS tags, which exist in similar form in most languages: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words).

Table 2 shows the two most common words for

[1] Pronouns and determiners are not explicitly annotated in the Russian treebank. As a result, the most common Russian nouns in the table are pronouns.

[2] On a standard benchmark (training on sections 1-18 of the Penn Treebank (Marcus et al., 1993) and testing on sections 22-24) our tagger achieves a state-of-the-art accuracy of 97.22%.



Raw Ngrams		Annotated Ngrams			
John	*short*	*_START_ John*	*John_NOUN*	*hair=>short*	*hair=>short_ADJ*
John has	*...*	*...*	*John has_VERB*	*hair=>black*	*...*
...	*short black hair*	*hair _END_*	*John _VERB_ short*	*_NOUN_<=has*	*_ROOT_=>has*

Figure 2: An English sentence and its part-of-speech tags and dependency parse tree. Below are some of the raw ngrams available in the first release of the Ngram Corpus, as well as some of the new, syntactically annotated ngrams.

some POS tag categories. It is interesting to see that there is overlap between the most frequent content words across language boundaries. In general, function words are more frequent than content words, resulting in somewhat less interesting examples for some POS tags. More typical examples might be *big* for adjectives, *quickly* for adverbs or *read* for verbs.

As suggested in Petrov et al. (2012), we train on the language-specific treebank POS tags, and then map the predicted tags to the universal tags. Table 3 shows POS tagging accuracies on the treebank evaluation sets using the 12 universal POS tags.

4.2 Syntactic Parsing

We use a dependency syntax representation, since it is intuitive to work with and can be predicted effectively. Additionally, dependency parse tree corpora exist for several languages, making the representation desirable from a practical standpoint. Dependency parse trees specify pairwise relationships between words in the same sentence. Directed arcs specify which words modify a given word (if any), or alternatively, which head word governs a given word (there can only be one). For example, in Figure 2, *hair* is the head of the modifier *short*.

We use a deterministic transition-based dependency parsing model (Nivre, 2008) with an arc-eager transition strategy. A linear kernel SVM with the following features is used for prediction: the part-of-speech tags of the first four words on the buffer and of the top two words on the stack; the word identities of the first two words on the buffer and of the top word on the stack; the word identity of the syntactic head of the top word on the stack (if available). All non-lexical feature conjunctions are included. For treebanks with non-projective trees we use the pseudo-projective parsing technique to transform the treebank into projective structures (Nivre and Nilsson, 2005). To standardize and simplify the dependency relations across languages we use unlabeled directed dependency arcs. Table 3 shows unlabeled attachment scores on the treebank evaluation sets with automatically predicted POS tags.

4.3 Syntactic Ngrams

As described above, we extract raw ngrams ($n \leq 5$) from the book text. Additionally, we provide ngrams annotated with POS tags and dependency relations.

The syntactic ngrams comprise words (e.g., *burnt*), POS-annotated words (e.g. *burnt_VERB*), and POS tags (e.g., *_VERB_*). All of these forms can be mixed freely in 1-, 2- and 3-grams (e.g., *the _ADJ_ toast_NOUN*). To limit the combinatorial explosion, we restrict the forms that can be mixed in 4- and 5-grams. Words and POS tags cab be mixed freely (e.g., *the house is _ADJ_*) and we also allow every word to be annotated (e.g., *the_DET house_NOUN is_VERB red_ADJ*). However, we do not allow annotated words to be mixed with other forms (e.g., both *the house_NOUN is _ADJ_* and *the house_NOUN is red* are not allowed). Head-modifier dependencies between pairs of words can be expressed similarly (we do not record chains of dependencies). Both the head and the modifier can take any of the forms described above. We use an arrow that points from the head word to the modifier word (e.g., *head=>modifier* or *modifier<=head*) to indicate a dependency relation. We use the designated *_ROOT_* for the root of the parse tree (e.g., *_ROOT_=>has*).

Language	POS Tags	Dependencies
English	97.9	90.1
Spanish	96.9	74.5
German	98.8	83.1
French	97.3	84.7
Italian	95.6	80.0
Russian	96.8	86.2
Chinese	92.6	73.2
Hebrew	91.3	76.2

Table 3: Part-of-speech and unlabeled dependency arc prediction accuracies on in-domain data. Accuracies on the out-of-domain book data are likely lower.

Figure 2 shows an English sentence, its POS tags and dependency parse tree, and some concrete examples of ngrams that are extracted. Note the flexibility and additional possibilities that the dependency relations provide. Using the raw ngrams it is not possible to accurately estimate how frequently *hair* is described as *short*, as there are often intervening words between the head and the modifier. Because dependency relations are independent of word order, we are able to calculate the frequency of both *hair=>black* and *hair=>short*.

Similarly, there are many ways to express that somebody is reading a book. The first plot in Figure 3 shows multiple related queries. The 3-gram *read _DET_ book* aggregates several more specific 3-grams like *read a book*, *read the book*, etc. The dependency representation *read=>book* is even more general, enforcing the requirement that the two words obey a specific syntactic configuration, but ignoring the number of words that appear in between.

5 Domain Adaptation

The results on the treebank evaluation sets need to be taken with caution, since performance often suffers when generalized to other domains. To get a better estimate of the POS tagging and parsing accuracies we conducted a detailed study for English. We chose English since it is the largest language in our corpus and because labeled treebank data for multiple domains is available. In addition to the WSJ (newswire) treebank (Marcus et al., 1993), we use: the Brown corpus (Francis and Kucera, 1979), which provides a balanced sample of text from the early 1960s; the QuestionBank (Judge et al., 2006), which consists entirely of questions; and the PPCMBE corpus (Kroch et al., 2010), which contains modern British English from 1700 to 1914 and is perhaps most close to our application domain.

	POS Tags		Dependencies	
Domain	base	adapted	base	adapted
Newswire	97.9	97.9	90.1	90.1
Brown	96.8	97.5	84.7	87.1
Questions	94.2	97.5	85.3	91.2
Historical	91.6	93.3	-	-

Table 4: English tagging and parsing accuracies on various domains for baseline and adapted models.

Since the English treebanks are in constituency format, we used the StanfordConverter (de Marneffe et al., 2006) to convert the parse trees to dependencies and ignored the arc labels. The dependency conversion was unfortunately not possible for the PPCMBE corpus since it uses a different set of constituency labels. The tagset of PPCMBE is also unique and cannot be mapped deterministically to the universal tagset. For example the string "one" has its own POS tag in PPCMBE, but is ambiguous in general – it can be used either as a number (NUM), noun (NOUN) or pronoun (PRON). We did our best to convert the tags as closely as possible, leaving tags that cannot be mapped untouched. Consequently, our evaluation results underestimate the accuracy of our tagger since it might correctly disambiguate certain words that are not disambiguated in the PPCMBE evaluation data.

Table 4 shows the accuracies on the different domains for our baseline and adapted models. The baseline model is trained only on newswire text and hence performs best on the newswire evaluation set. Our final model is adapted in two ways. First, we add the the Brown corpus and QuestionBank to the training data. Second, and more importantly, we estimate word cluster features on the books data and use them as features in the POS tagger.

The word cluster features group words deterministically into clusters that have similar distributional properties. When the model encounters a word that was never seen during training, the clusters allow the model to relate it to other, potentially known words. This approach improves the accuracy on rare words, and also makes our models robust to scanning er-

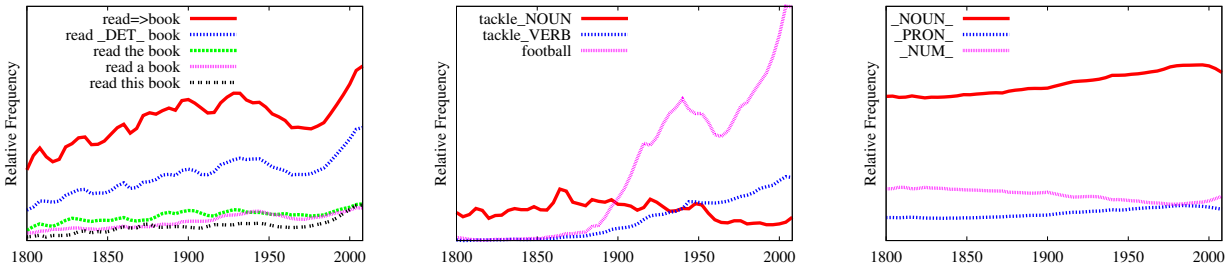


Figure 3: Several queries expressing that somebody is reading a book (left). Frequencies of *tackle* used as noun vs. verb compared to the frequency of *football* (middle). Relative frequencies of all nouns, pronouns and numbers (right).

rors. For example, in older books the medial-s ($\int$) is often incorrectly recognized as an 'f' by the OCR software (e.g., "beft" instead of "best"). Such systematic scanning errors will produce spurious words that have very similar co-occurrence patterns as the correct spelling of the word. In fact, a manual examination reveals that words with systematic scanning errors tend to be in the same cluster as their correctly spelled versions. The cluster feature thus provides a strong signal for determining the correct POS tag.

While the final annotations are by no means perfect, we expect that in aggregate they are accurate enough to be useful when analyzing broad trends in the evolution of grammar.

6 Conclusions

We described a new edition of the Google Books Ngram Corpus that provides syntactically annotated ngrams for eight languages. The data is available for download and viewable through an interactive web application at `http://books.google.com/ngrams`. We discussed the statistical models used to produce the syntactic annotations and how they were adapted to handle historical text more robustly, resulting in significantly improved annotation quality. Analyzing the resulting data is beyond the scope of this paper, but we show some example plots in Figure 3.

References

M. Davies. 2011. Google Books (American English) Corpus (155 billion words, 1810-2009). In *http://googlebooks.byu.edu/*.

M.-C. de Marneffe, B. MacCartney, and C. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proc. of LREC*.

W. N. Francis and H. Kucera. 1979. Manual of information to accompany a standard corpus of present-day edited American English. Technical report, Brown University.

J. Judge, A. Cahill, and J. van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proc. of ACL*.

A. Kroch, B. Santorini, and A. Diertani. 2010. Penn parsed corpus of modern british english. Technical report, LDC.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.

E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M. A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature*.

C. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? *Proc. of CICLing*.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19.

J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*.

J. Nivre and J. Nilsson. 2005. Pseudo-projective dependency parsing. In *Proc. of ACL*.

J. Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

S. Petrov, D. Das, and R. McDonald. 2012. A universal part-of-speech tagset. In *Proc. of LREC*.

M. Ravallion. 2011. The two poverty enlightenments: Historical insights from digitized books spanning three centuries. *Poverty And Public Policy*.

J. Uszkoreit and T. Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proc. of ACL*.