Corpora of Typical Sentences

Lydia Müller, Uwe Quasthoff, Maciej Sumalvico
University of Leipzig
Augustusplatz 10, 04109 Leipzig
{lydia,quasthoff,sumalvico}@informatik.uni-leipzig.de

Abstract

Typical sentences of characteristic syntactic structures can be used for language understanding tasks like finding typical slotfiller for verbs. The paper describes the selection of such typical sentences representing usually about 5% of the original corpus. The sentences are selected by the frequency of the corresponding POS tag sequence together with an entropy theshold, and the selection method is shown to work language independently. Entropy measuring the distribution of words in a given position turns out to identify larger sets of near-duplicate sentences, not considered typical. A statistical comparison of those subcorpora with the underlying corpus shows the intended shorter sentence length, but also a decrease of word frequencies for function words associated to more complex sentences.

Keywords: typical sentences, sentence signatures, language statistics, corpus comparison

1. Introduction

Statistical analyses of language are usually based on large corpora compiled from publicly available written sources, e.g. news, Wikipedia, crawled webpages or literature (Baroni and Bernardini 2004, Biemann et al. 2013). Compared to everyday speech, such sources tend to be biased towards long sentences and complex syntactic structures. There have been attempts to compile corpora more represantative of everyday language by utilizing different sources, especially movie subtitles (Lison and Tiedemann 2016). However, the availability of such sources is limited.

In this paper, we propose a method for diminishing the bias towards complex syntactic structures usually found in large corpora. In order to accomplish this, we select "typical sentences", defined as sentences with a common syntactic structure (represented as a sequence of POS-tags). In contrast to simplified or controlled languages (like Simplified English (Ogden 1932) or Kontrolliertes Deutsch (Lehrndorfer 1996) (controlled German) there is no set of handwritten rules for syntax and vocabulary. However, the sublanguage emerging in typical sentences may be considered as an automatically genarated analogy.

As an illustration of our approach, we currently provide typical sentence corpora for English, German, French, Dutch and Italian as a part of the Leipzig Corpora Collection.[1] Random samples of up to 1 million sentences are freely available for download.

2. Selecting typical sentences

Given a POS-tagged sentence, we define the *sentence signature* as the corresponding sequence of POS tags. For a corpus of sentences, such signatures can be ordered by frequency. As one could expect, the most frequent sentence signatures belong to relatively short sentences with typical structure. Table 1 shows the top five sentence signatures from newspaper corpora of the Leipzig Corpora Collection (Goldhahn et al. 2012) with sample sentences for English. As POS tagger, the Stuttgart Tree Tagger (Schmid 1994) is used.

The following two properties of the list of sentence signatures are not so obvious: There are frequent sentence signatures belonging to large sets of near-duplicate sentences. They are usually unexpectedly long and contain numbers. The most frequent examples are given in Table 2. Such sentences are not considered typical and should be identified and removed. A simple way to identify such near-duplicates is to exploit the small variation in the vocabulary for these sets of sentences: For each sentence signature (with minimum frequency 5), the normed entropy[2] of the vocabulary distribution in each position of the sentence is calculated. Figure 1 shows the distribution of the median normed entropy for sentence signatures for German. Distributions for other languages are similar. Sentence signatures with low median normed entropy (≤ 0.5) are removed.

The choice of the median rather than arithmetic mean is motivated as follows: it is acceptable and expected, that some positions in the sentence show little or no variation in the vocabulary (e.g. articles, auxiliary verbs, even the main verb). As long as enough other positions show considerable variation (e.g. subjects and objects), the signature is a good candidate for a source of typical sentences. The median enables us to separate the signatures with variation at too few positions more clearly.

Another interesting observation is that there are unexpectedly many different sentence signatures: about 95.6% of all sentences signatures appear only once. Table 3 shows the sentence coverage for the N most frequent sentence signatures after the removal of the near-duplicates.

After removing signatures corresponding to near-duplicate sentences, we select $100,000$ most frequent signatures to form the typical sentence corpora. Although the choice of this theshold is quite arbitrary, it leads to a good tradeoff between the corpus size and the simplicity of sentences by selecting typically between 5 and 10 percent sentences (see Table 4 for details).

3. Statistical properties

In order to compare the subcorpus of typical sentences to the original corpus, we consider the following properties:

[1] http://corpora.uni-leipzig.de

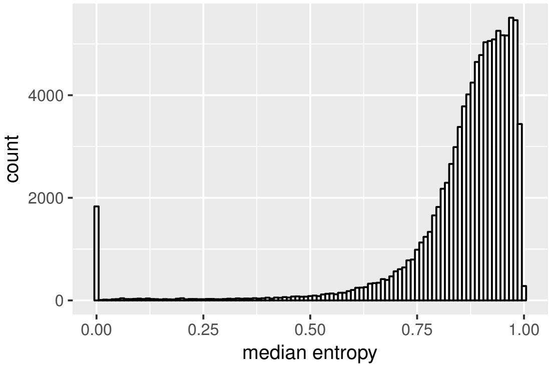
[2] The entropy is divided by the maximum entropy possible at a given position, so that the result is a number between 0 and 1.

Rank	Signature	Sample sentence
1	DT NN VBZ JJ SENT	The future is mobile.
2	PP VBD RB JJ SENT	It was just crazy!
3	DT NN VBZ RB JJ SENT	The news is all good.
4	PP VVD DT NN SENT	They licensed the technology.
5	PP VVP DT NN SENT	They love the area.

Table 1: Top 5 signatures for English with example sentences.

Signature	Sample sentence
NP NP VVD TO DT NN SENT	Aaron Blake contributed to this report.
NP NP VVD IN CD NN IN NP , NP CD , CD SENT	Alma Santos posted at 7:30 am on Sun, Oct 11, 2015.
DT NN NN IN DT NN VBZ (CD) CD SENT	The phone number at the clinic is (320) 395-2527.

Table 2: Sample signatures corresponding to near-duplicate sentences in English.



O

Figure 1: Normed entropy distribution for German sentence signatures: For each signature, the median of position-wise normed entropies is calculated.

N	% sentences
10,000	4.1 %
50,000	6.8 %
100,000	8.1 %
500,000	11.7 %
1,000,000	13.4 %
5,000,000	18.5 %
10,000,000	22.1 %

Table 3: Sentence coverage of the N most frequent signatures in the German corpus (total=$211,657,876$).

- number of sentences (see Table 4)
- distribution of sentence length (see Figure 2)
- Zipf distribution for word frequencies, frequency@200,000
- Changes in the stopword ranking
- Number of significant word co-occurrences

The described selection of sentences leads to a strong reduction of sentences in the corpora with typical sentences. For example, the German corpus with typical sentences contains only 7.76% of the all German sentences. However, this are still more than 20 million sentences (see Table 4). Thus, the corpora are still large enough for statistical analyses.

Table 4: Number of sentences

Language	sentences	typical sentences	percent
German	259,026,023	20,103,234	7.76
English	156,934,303	4,193,396	2.67
French	74,823,426	3,758,924	5.02
Dutch	70,332,253	5,432,161	7.72
Italian	44,636,533	2,023,640	4.53

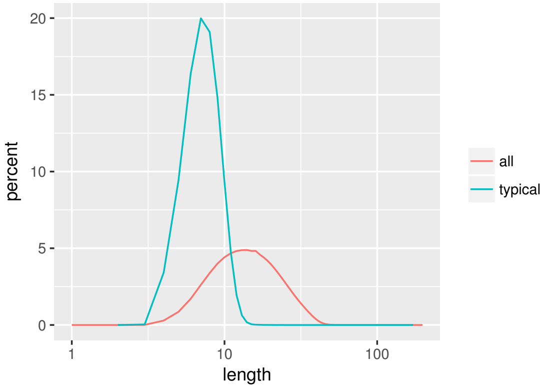


Figure 2: Sentence Length Distribution: Sentence length is measured as number of tokens. Red and turquoise line show the distribution for all sentences and only the typical sentences of German, respectively.

We measure the sentence length as the number of tokens in a sentence. Typical German sentences are normally about 7 tokens in length, while the peak in the distribution for all German sentences is 14 tokens (see Figure 2).

Zipf's Law for the German corpora is shown in Figure 3. The selection of only sentences with frequent signatures does not affect the characteristics of the distribution. Due to the smaller corpus size of the typical sentences, the distribution is shifted to lower frequencies.

The simpler structure of shorter sentences results in

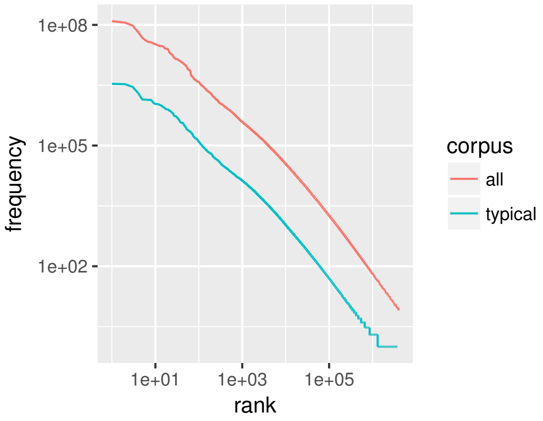


Figure 3: Zipf's Law for typical and all sentences of German.

word	type	rank	rank change
Die	article	3	–12
Das	article	6	–33
Der	article	7	–20
werden	auxiliary	13	–11
Sie	pronoun	14	–23
und	conjunction	15	+12
von	preposition	17	+11
sind	auxiliary	22	–11
wurde	auxiliary	26	–28
im	preposition	28	+15
war	auxiliary	30	–19
zu	preposition	31	+24
Es	pronoun	33	–33
Ein	article	36	–54
bei	preposition	39	+11
Er	pronoun	43	–54
In	preposition	45	–16
Eine	article	46	–83
am	preposition	47	+15
nach	preposition	49	+18
Wir	pronoun	50	–39

Table 5: Rank changes among the top-100 German words.

changes in the stopword ranking (Table 5 and 6). As could be expected, more sentences begin with simple nominal phrases and so the ranks of capitalized articles and pronouns decrease. Also the ranks of copula/auxiliary verbs decrease, as they play a key role in typical sentences. More interesting is an increase for conjunctions, especially "and", as well as prepositions. Conjunctions automatically generate longer sentences and are variable in its syntactic position. Also prepositional phrases often give rise to complex syntactic structures. These features are less frequent in typical sentences.

The number of significant co-occurrences depends on the number of occurrences of a word and is therefore not comparable across corpora of different sizes or across large sets of words of different frequencies. In Figure 4, we use the neighbors of a word as input for the calculation of significant co-occurrences. Thus, a word occurring n times co-occurs at most $2n$ times with one of its significant co-occurrences (i.e. a significant co-occurrence on the left and the right side). Therefore, we normalize the frequency of the significant co-occurrences for each word by $2n$. The distributions of the resulting ratios is shown Figure 4. The quality of word co-occurrences also changes: Due to the decrease of conjunctions, the number of pairs of similar terms decreases. Hence, there are fewer similar terms in the sentence-based word co-occurrences.

word	type	rank	rank change
is	auxiliary	3	–5
The	article	4	–7
was	auxiliary	5	–8
I	pronoun	9	–9
We	pronoun	12	–48
He	pronoun	13	–46
and	conjunction	14	+11
It	pronoun	16	–46
They	pronoun	20	–98
at	preposition	23	+9
can	modal	37	–11
from	preposition	38	+16
A	article	42	–40

Table 6: Rank changes among the top-100 English words.

It is notable that after selecting for typical sentences, words more often appear with significant co-occurrences than before. This indicates a lower variety in the combination of words when using only typical sentences. Interestingly, there is a peak in the distribution for typical sentences at 0.5. These are words that always co-occur with a significant co-occurrence.

Figure 4 shows two distributions for each corpus of German. Hereby, the solid lines take only frequent words (frequency $\geq$ 100) into account, while the dashed lines show the distribution for all words. Since the corpus with typical sentences is much smaller, the word frequencies are lower and more noisy. This leads to artificial peaks in the distribution. However, the general trend is the same no matter if we use all words or only frequent words.

4. Application: Clustering of sentences and typical constituents

For high and medium frequency verbs, we find many typical sentences with the same verb in the same position. By means of clustering those sentences using word similarity, we can identify possible replacements for words in each position. For the German verb *brannte* (engl. *burned*) we find the following typical sentence of length five: *Die Scheune brannte völlig nieder. (The barn burned down completely.)* Table 7 shows the replacement words in similar sentences, while Table 8 shows an equivalent example from the English corpus. In the German example, it turns out that both buildings (or its parts) and vehicles are mentioned to burn completely. So, in position two the burning objects built two clusters, in positions four and five we find near-synonyms for formulating the sentences.

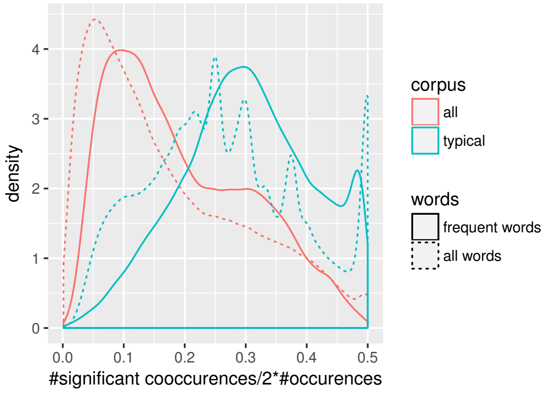


Figure 4: Co-occurrences in typical and all sentences for German. We calculate the ratio of the observed significant co-occurrences and the theoretically possible co-occurrences (i.e. 2 times the occurrence of a word). The figure shows the density distribution of these ratios for the German corpus of all sentences and the German corpus of typical sentences. Solid lines represent the distribution taking only frequent words (occurrences ≥ 100) into account, while dashed lines correspond to the distribution for all words.

Table 7: Sentence variants for German

Pos.	Sentence	Variants
1	Die	Der, Ein
2	Scheune	Gebäude, Haus, Halle, Dachgeschoss, Auto, Fahrzeug, Wagen ...
3	brannte	
4	völlig	vollständig, vollkommen, komplett, total
5	nieder	ab, aus

5. Language Resource References

In the Leipzig Corpora Collection, POS tagging is applied for about 35 languages with the following 21 languages having more than 10 million of sentences: Arabic, Danish, Dutch, English, Esperanto, French, German, Hungarian, Icelandic, Italian, Norwegian, Polish, Portugese, Romanian, Russian, Slovak, Slovene, Spanish, Swedish, Ukrainian and Vietnamese. Here, mainly the Stuttgart Tree Tagger (Schmid 1994) is used.

For these languages, the corpora of typical sentences will subsequently be produced and made available for download at *http://wortschatz.uni-leipzig.de/en/download*.

At the time of writing this paper, the corpora for English, French, German, Dutch and Italian are already available. The corpora can be recognized on the download website by the '`-typical`' addition in their names.

References

Marco Baroni and Silvia Bernardini. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316, 2004.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. Scalable construction of high-quality web corpora. *JLCL*, 28(2):23–59, Dec 2013.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

A. Lehrndorfer. *Kontrolliertes Deutsch: linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der technischen Dokumentation*. Tübinger Beiträge zur Linguistik. G. Narr, 1996.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, 2016.

C.K. Ogden. *Basic English: a general introduction with rules and grammar*. Psyche miniatures: General series. K. Paul, Trench, Trubner & Co., Ltd., 1932.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.

Table 8: Sentence variants for English

Pos.	Sentence	Variants
1	The	
2	whole	–
3	building	schoolhouse, village, house
4	was	had, –
5	completely	all, –
6	burned	
7	down	out