Evaluation of the Hypotheses

Hypotheses

In this part, we will first discuss each hypothesis from the Hypotheses pdf. Note that for the sake of an overview, we did not list the hypotheses again.

1.1. This hypothesis was rejected, the most common word is very clearly “ist” and that is independent of whether it is sb or an object. This makes sense because trees are often described as being like something, like a fact.
1.2. Similarly to 1.1., the most common word found with this construction is “ist”. While we did predict a verb, we predicted it to be a more active verb, that would take an object. We did not expect it to be “sein”.
1.3. The word ‘Baum’ occurred about twice as many times in news articles as in Wikipedia, so that we can confirm this hypothesis.
1.4. As discussed in 1.3., ‘Baum’ occurs more often in news articles, thus, Wikipedia has fewer mentions.
1.5.The word ‘Baum’ occurs about more or less the same in both eras, which goes against our intuition that it would occur more often now. Since ‘Baum’ belongs to a very basic kind of vocabulary though, this makes sense.
2.1. Our expectation held true, that ‘ist’ will be among the most common words. Interestingly, articles on the web will use ‘war’ more often than ‘ist’.
2.2. While our prediction was deemed incorrect, there are some issues with the result. The most common word is once again a variation of ‘sein’. However, nk usually requires the adjective to be declined in German. Since spacy does not allow to work with .tsv files, we were not able to search for lemmas. Thus, this shows a very inaccurate picture of the real situation. These data points, in our opinion, should not be taken into account.
2.3. We found that ‘schön’ has very similiar frequency levels in both news articles and web pages. Very interestingly though, the most common partner word in news articles is in present tense, where the most common partner word on web pages is in past tense. This makes sense, since newspapers tend to write about current events, where web pages appear to have more freedom in that regard.
2.4. Against our intuition, the time period does not seem to make a difference in the frequency of the usage of ‘schön’. This would make us believe that beauty is as much of importance now, as it was a decade ago.
3.1. In most cases, the partner word is ‘es’, which is a pronoun. While this goes against our predictions, it makes sense. ‘Es’ can be used for general statements, such as: ‘Es ist gut’, without having to specify what is being talked about.
3.2. We were right, that the most commonly used partner words were adjectives, namely ‘unklar’ and ‘gemein’. However, we could not have predicted the actual adjectives that occurred the most frequently.
3.3. ‘Sein’ occurs very frequently in both text types, which was to be expected. Its wide usage makes it one of the most used verbs imaginable.
3.4. Results here are very similar to 3.3. The explanations are the same as in 3.3.

To sum up, our intuition about what partner words would occur together most often was mostly off. We suspect that this prediction is very data or corpus dependent. Many of the hypotheses about time periods and text types were correctly assumed, which confirms our initial intuitions.

Intuition

The raw frequencies appear to be much more intuitive for the end user, since it allows to estimate the importance of the partner word. However, for machines, the LogDice value is much more useful, since it portrays an objective value, which can be used regardless of the corpus' size. Overall, the raw frequency is a value which is easier to visualize and imaginable - in the head - compared to the LogDice value, for which we first had to understand what this word means by reading a paper and implementing the formula.

Surprises + Weakness

It was surprising, that spacy does not know how to deal with tsv files. This was unfortunate since it did not allow for lemma searches within the input sentences. This leads to only being able to search the verbatim spelling of the words given by our hypothesis. Furthermore, the way that the dependencies were structured was changed in one script, which is something we would not do again. In general, we were impressed by how many hours we invested in this assignment.

Lemmatizer

The lemmatizer was surprisingly good, finding most words correctly. We therefore did not customize by extending the search of lemmas in the patterns - as suggested in the assignment.The bigger issue was the dependencies. Often they were obviously wrong. 'Der Baum steht' would sometimes be marked as object, which neither grammatically nor semantically makes sense.