

# **PhD course in Statistical Methods**

## **Project / term paper**

**Spring 2009**

**Maria Eskevich**

Comparison of the -n-grams for language modeling for four languages

(English, German, French and Russian)

### **Introduction**

The author takes part in the development of the speech and language recognition systems. In order to fulfill these tasks we need language models for each language. The most common way of completing this task is to build -n-gram language models. Therefore, this project consists of building -n-grams (unigram, bigram, trigram) and comparing them.

### **Material/Method**

The material used for this experiment is a multilanguage corpus consisting of news texts built on the basis of Internet sources (with the volume of the same amount for each language - more than 400 000 documents).

Before building the -n-grams models the texts are normalized. It includes such operations as automatic correction of misprints, automatic modification of the word register. For example, if the word is not presented in the vocabulary of proper names the modification of its register is made according to the ratio of different spellings, only one way of spelling is chosen.

The other important type of normalization is the conversion of numerated numbers into words. There is no unique way of doing it in all languages because of their different grammatical structures. In English and French there are no cases, therefore the conversion from numbers is quite direct, we have

only to distinguish the type of the numeral – where it is ordinal or cardinal. Usually it is obvious from the text, because the former have endings (e.g. 1<sup>st</sup>, 1er). German and Russian have cases and the gender of dependent nouns affects the spelling. Since the deep part of speech analysis is not applied at the moment these numbers were just replaced by special tags <UNK-digital>. We can not simply omit these cases, since they compose about 5 % of the word usages and the resulting bi- or trigrams would be mistaken.

When we have prepared clear texts we build the frequency vocabulary. On their basis we create the n-gram files.

We load the text cutting it into words and looking for these words in the sorted frequency vocabulary. For every 2 and 3 sequenced words we build a bi- or trigram correspondingly. If this n-gram is already presented in the vocabulary we just increase its frequency, otherwise we just add it to the list. In the end we sort the resulting n-grams by frequency and cut from the list the n-grams with low frequencies (equal to 1).

## Results

Figure 1 shows the overall number of n-grams for 4 languages. It can be seen that the number of bigrams for Russian is twice bigger than for other languages. It is caused by the inflectional morphology of the language and free order in the sentences. However, the number of trigrams for Russian is the minimum among the languages. It is caused by the sampling of n-grams.

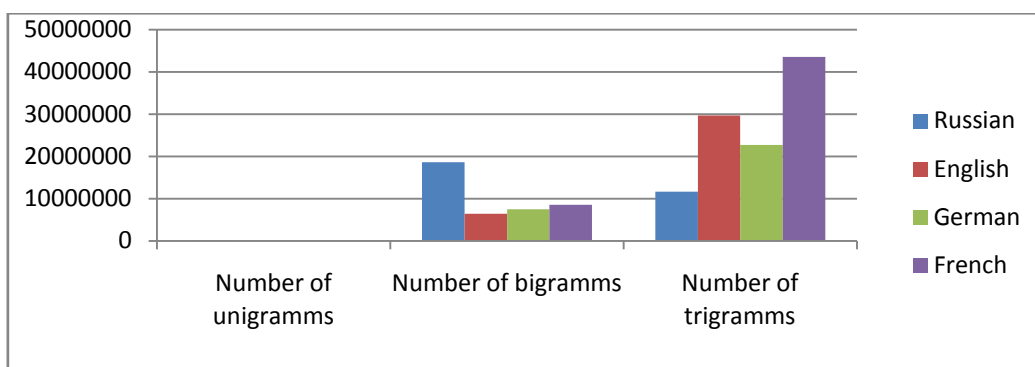


Figure 1. Number of n-grams for 4 languages

This figure might give rise to 2 hypothesis: either the corpus of French is more uniform in the lexicon (indeed, the number of issues for French is three times less than for Russian or two times less than for German) or/and we need more data for Russian to build the comparable number of trigrams (or we have to change the strategy of n-gram building for Russian, probably not building them based on words, but on stems).

Figure 2 shows separately the number of unigrams. The tendency of more inflectional languages towards having more units is consistent. The number of unigrams for German is even higher than for Russian. It is probably caused by the fact that there are more regional issues in German and therefore the proper names are more varied.

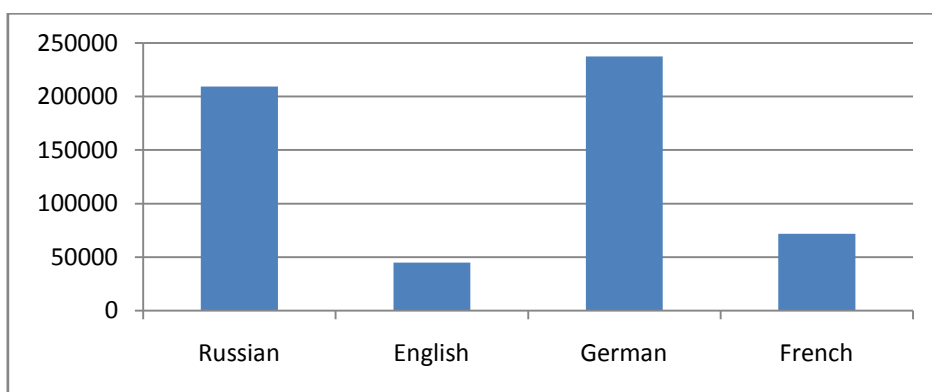


Figure 2. Number of unigrams for 4 languages

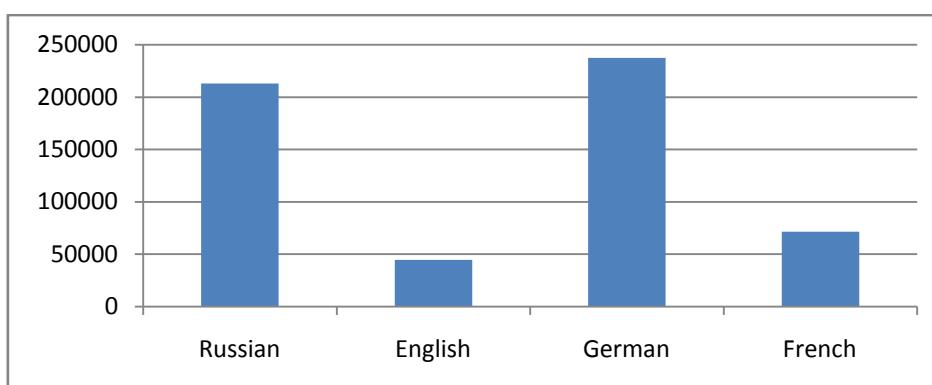


Figure 3. Volume of the vocabulary for covering 98 % of the text.

Figure 3 also confirms that morphologically less diverse language, English, is covering 98 % of the text with the least number of words.

## **Conclusion**

The comparison of n-gram lists for languages with different morphological and syntactic structures has shown that this method requires specification for groups of languages (highly inflectional and more analytic languages, the additional parameter here can be the degree of word order freedom). This might be the better way for creating multilanguage corpora with less distortion in data for different languages.