

# Measuring Monolinguality

Uwe Quasthoff, Chris Biemann

NLP Department,  
Faculty of Mathematics and Computer Science  
University of Leipzig, Germany  
{quasthoff,biem}@informatik.uni-leipzig.de

## Abstract

We present an approach to measuring the amount of material in a natural language text corpus that consists of text in languages other than the main language. Having a presumably monolingual corpus at hand, we ask for the amount of multilingual noise by comparing the frequency of high-frequent words in monolingual corpora of different languages to their frequency in the corpus in question. The ratio of the expected and the measured frequencies per language quantifies the amount of noise per language. The measure is very effective since it requires only the comparison of a few thousand frequency counts.

We evaluate the method by artificial mixtures of two language corpora for different noise levels and demonstrate the effect of a corpus cleaning method by measuring monolinguality before and after cleaning.

## 1. Introduction

When building a large corpus for a given language, one has to assure that the data are as clean as possible. This is especially important when using resources from the Web, where neither top-level-domain nor source guarantees monolinguality in any respect. See (Kilgarriff, 2001) for a discussion about the web as a corpus; multilinguality on the same web servers is even employed by (Resnik and Smith 1995) to construct aligned bilingual corpora. The definition for cleanliness may vary according to the principles chosen for the corpus construction. For this paper, we want to assume the following:

- The corpus is sentence separated and the order of the sentences is not important.
- Clean means monolingual, i.e. during pre-processing, sentences belonging not to the specified language should be identified and removed.

The question whether a sentence belongs to a given language is not at all trivial because a German sentence can, for instance, contain an English movie title or a Latin medical term. The principles for corpus construction might contain hints how many foreign language objects should be tolerated. Usually one will allow such isolated foreign language items, but not foreign language sentences (which might contain some words in the corpus language). At this tolerance level (i.e. we allow only a few foreign language objects), the cleaning becomes more difficult and one has to check the quality of the cleaning process.

The aim of the paper is not to describe algorithms for cleaning procedures but to measure their result. A numerical value of monolinguality can be considered as a quality measure for corpora. This measure should also give satisfactory results if the languages considered have some words in common.

Note that the usual tests for language detection (e.g. described in (Dunning, 1994)) are not sufficient for cleaning because they allow a considerable amount of multilingual material to pass, dependent on document length. The test described here should be able to quantify this amount. As shown in the examples, foreign language

material of 0.001% can be measured. The lower bound of the verifiability depends only on the corpus size.

The availability of clean monolingual resources is important for a variety of applications. To name a few, methods that construct language models from corpora (e.g. Brown et al. 1992) will be disturbed by alien language material and morphology induction (like described in (Goldsmith, 2001) inter al.) will face undesired problems. In dimensionality reduction steps (e.g. Derweester et al. 1990), some of the dimensions will be occupied by other languages than the target language, hampering performance.

## 2. A Measure for Monolinguality

### 2.1. Informal description

We propose a measure, which distinguishes between random foreign noise, and foreign language objects of a certain special kind like proper names, quotations etc. While the latter might be allowed in a corpus of language A, we will measure mainly typical text of another language B contained in the corpus. Such typical text will contain nearly all high frequency words of language B.

If the absolute amount of word of language B is large enough, their distribution will be like in an ordinary language B corpus for many of these words. Hence, many of those words will be a similar ratio of their usual relative frequency compared to their relative frequency in the language A corpus.

Of course, this is not true for all words of language B under consideration. Exceptions are words often used in typical foreign language objects like named entities or titles. And, of course in the case of words being used in both languages A and B. However, their number turns out to be surprisingly low.

Hence, we get a clear peak when counting the number of words for different frequency ratios. Moreover, the resulting peak does not depend on the number of high frequent words used. In the examples, we use always the 1000 most frequent words.

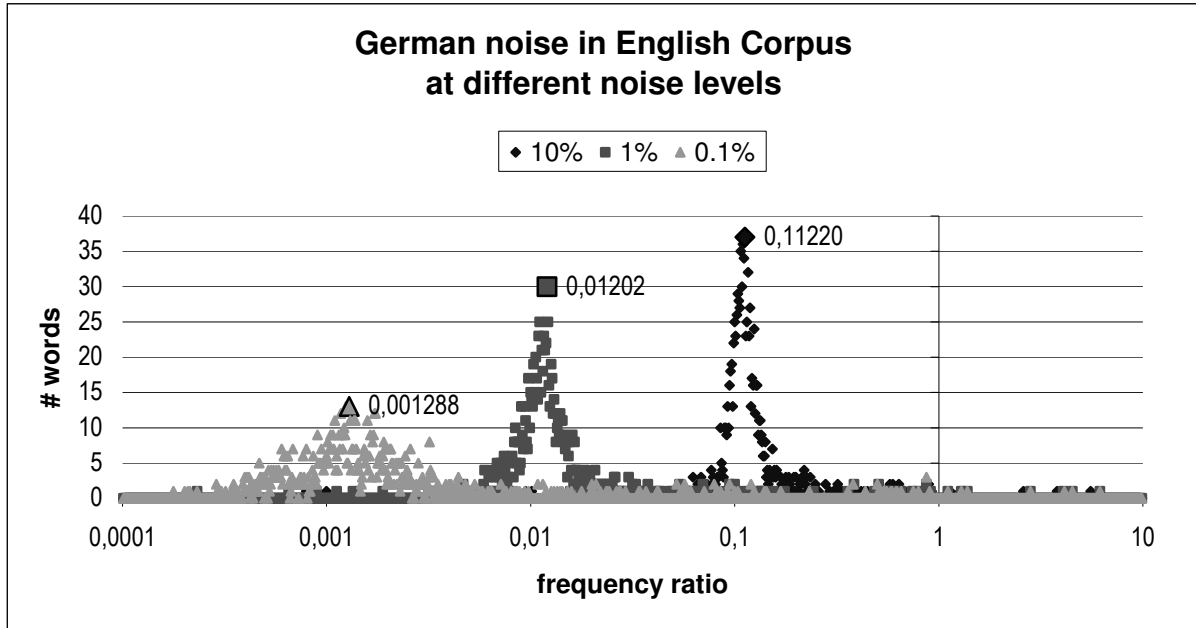


Figure 1: German noise in English corpus. The numbers attached to the peaks are the results of our measure (experiment 1a)

## 2.2. Comparing high frequency words

Assume a corpus of a language A contains  $x\%$  of noise of some language B. Moreover, the corpus should be large (say, more than 1.000.000 sentences) and the noise should be typical text of language B. Then we consider the top-1000 high frequent words of language B. If such a high frequent word  $w$  is not contained in language A, it should appear in the corpus with a relative frequency of roughly  $x\%$  of its relative frequency in language B. If  $w$  is also a valid word in language A, its relative frequency will be much higher. We define the frequency ratio of  $w$  as the relative frequency of a word  $w$  in A divided by its relative frequency in the corpus B.

There are four groups of words in the top 1000 words of language B:

- Words that do not occur in language A. Their frequency ratio will be around  $x\%$ .
- Words that are also amongst the highest frequency words of language A and moreover have the same function. Their frequency ratio will be around 1.
- Words that occur in language A, but at different frequency bands. They are a random sample of words of L and distributed in a Zipf way, cf. (Zipf, 1949).
- Words of B that are often used in named entities and titles (such as capitalized stop words). They appear in the corpus of language A more frequently than the expected  $x\%$  of noise.

The second group of words is only present in languages that are very similar to each other. Table 1 shows overlaps in the top 1000 words of some European languages.

	da	de	ee	en	es	fr	is	it	nl	no
de	36									
ee	11	5								
en	41	26	11							
es	18	14	7	27						
fr	33	19	10	59	52					
is	43	13	7	9	6	11				
it	31	11	9	25	98	51	9			
nl	69	56	10	52	25	40	21	30		
no	489	33	18	38	25	35	55	40	64	
se	221	23	15	27	23	32	50	32	54	257

Table 1: overlap in some European languages with regard to the most frequent 1000 words: Danish (da), German (de), Estonian (ee), English (en), Spanish (es), French (fr), Icelandic (is), Italian (it), Dutch (nl), Norwegian bokmål (no), Swedish (se)

## 2.3. The dominant frequency ratios

In the figures 1 and 2, we have the frequency ratios at the x-axis ranging from  $10^{-4}$  to 10 on a logarithmic scale. After discretizing the frequency ratios it is counted, how many words fall into the corresponding intervals. We find a Gaussian shaped curve with a clear maximum at the amount of noise at  $x\%$  caused by words of group 1, a similar peak near 1 due to the second group (if the languages are similar) and some uniformly distributed noise introduced by the words of group three. Words of group four are scattered between  $x\%$  and 1.

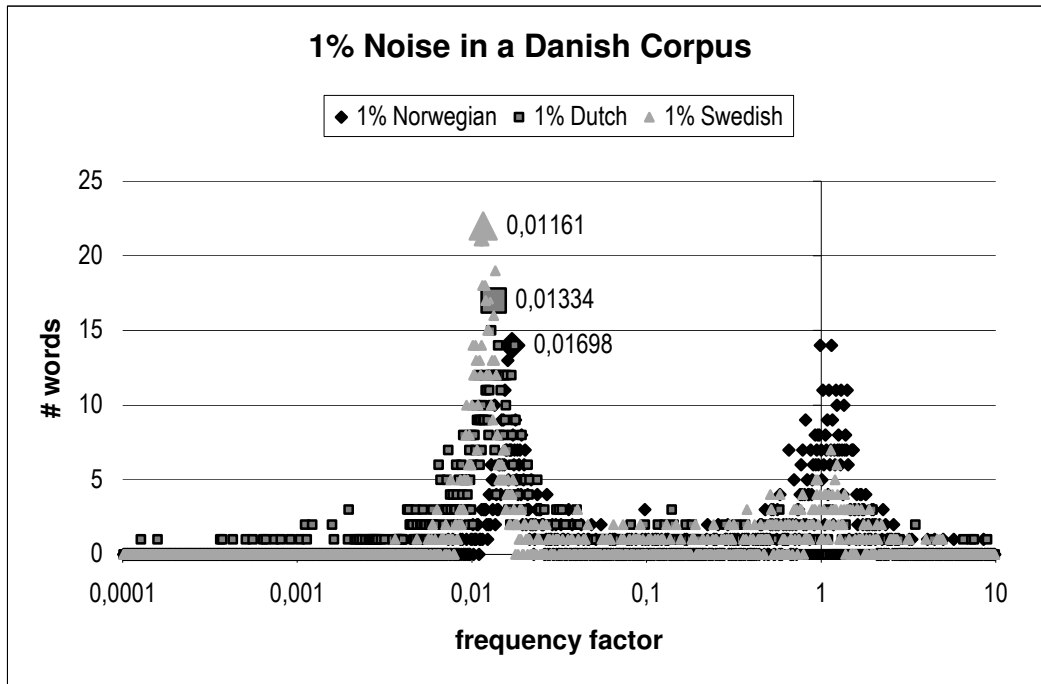


Figure 2: Norwegian, Dutch and Swedish noise in Danish corpus (experiment 1b)

### 3. Experiments

In the following two examples we search for foreign language noise in English and German corpora. In the first experiment, this noise is manually inserted into the British National Corpus (BNC, <http://www.natcorp.ox.ac.uk/>, (Leech, 1992)). As a result we should measure the exact amount of the previously inserted noise. Moreover, the effect of very similar languages is discussed using Scandinavian languages.

The second experiment uses randomly collected text from the web using only .de-domains. For the reduction to a corpus in German language, foreign language sentences are automatically removed. The amount of foreign language text is shown before and after cleaning.

#### 3.1. Experiment 1: Artificial Noise

In order to test our measure we performed two experiments with introducing noise in monolingual corpora. In experiment 1a we aimed at finding out how well the measure captures different noise levels, in experiment 1b we tested very similar languages.

Figure 1 shows the frequency ratio interval counts for a 10%, 1% and 0.1% German noise as taken from <http://www.wortschatz.uni-leipzig.de> (Biemann et al., 2004) injected in a chunk of the BNC corpus. All mixtures consisted of about 20 Million tokens. The noise levels measured are slightly larger than expected, see figure 1. This is due to the fact that the BNC corpus contains German sentences (some containing errors) like e.g.

- Geschichte in Literatur und Film seit den sechziger Jahre , in : Geschichte als Literatur , ed .
- Cantatas No. 140 , Wachtet auf , ruft uns die Stimme ; No. 147 , Herz und Mund and Tat und Leben .

- Prince : Hans Adam von und zu Liechtenstein II .
- Nur an den beiden Poien menschlicher Verbindung , dort , wo es noch keine oder keine Worte mehr gibt , im Blick und in der Umarmung , ist eigentlich Glück zu finden , denn nur dort ist Unbedingtheit , Freiheit , Geheimnis und tiefe Rücksichtslosigkeit .

Figure 2 depicts the distribution for very similar languages (in terms of table 1). Again, the measure deviates not severely from the goal of 1% noise. Material was taken from <http://corpora.informatik.uni-leipzig.de> to build corpora of about 17 Million words.

#### 3.2. Experiment 2: Web Text

Experiment 2 uses a corpus of about 40 million sentences randomly collected from .de-domains. We measured the amount of foreign languages before and after cleaning, which was carried out as outlined in (Quasthoff et al. 2006). Table 2 contains not only the main frequency ratios, but also the number of top-1000-words of a foreign language found in the corpus. As stated in 2.1,

	Before cleaning		After cleaning	
	Number of top-1000-words found	Approx. Frequency ratio	Number of top-1000-words found	Frequency ratio
German	1000	0.708	1000	0.946
English	995	0.126	987	0.0010
French	924	0.0398	906	0.00002
Dutch	995	0.000891	775	0.000006
Turkish	642	0.0000631	562	0.000006

Table 2: frequency ratios and number of top 1000 words when cleaning a German web corpus

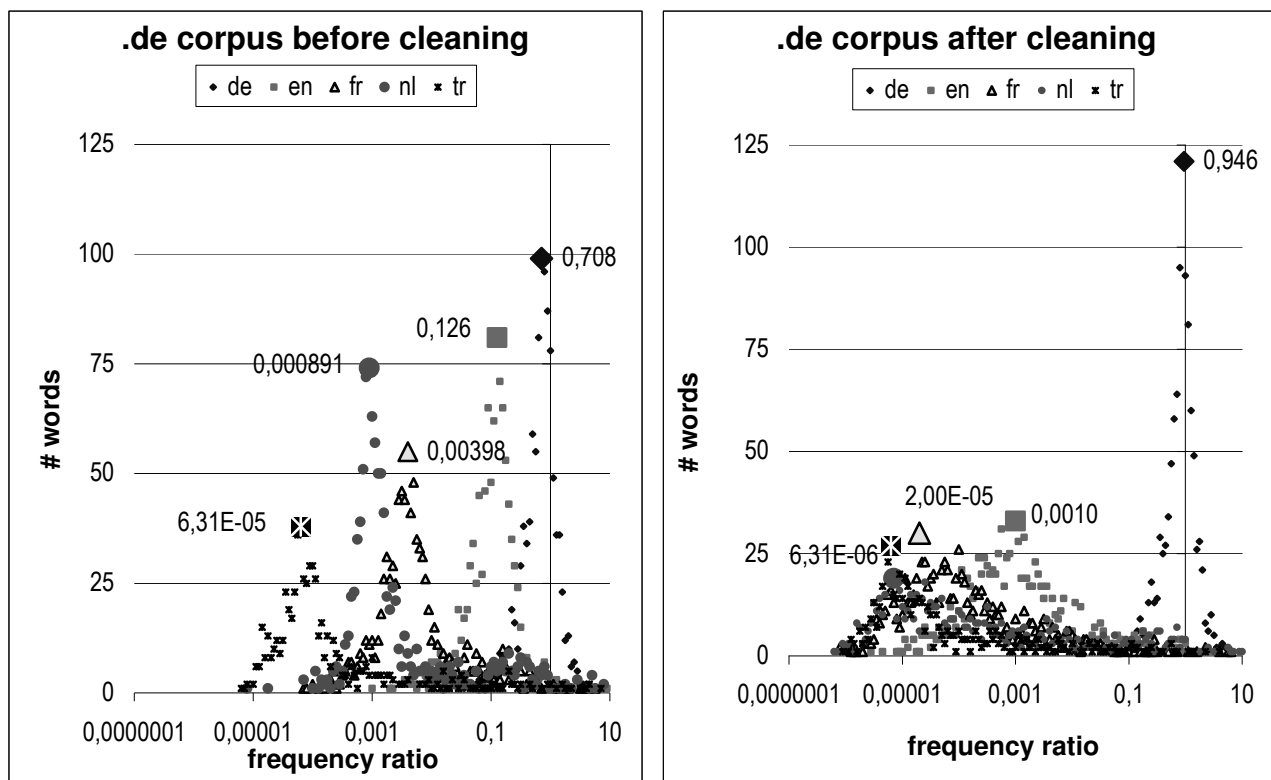


Figure 3: The effects of corpus cleaning with regard to the monolinguality measure

(nearly) all of the top-1000-words of the noise language are expected to appear in the corpus. The smaller numbers for Turkish and Danish (only in the cleaned version) indicate that the limits of the method are reached in the case where (size of top-wordlist) / (frequency ratio) has the same order of magnitude as the frequency of the most frequent word in the corpus. Figure 3 visualizes the findings of table two.

Moreover, the table shows that language cleaning can reduce the noise by a factor of at least 100. The corresponding noise can be measured down to a frequency ratio of approximately  $10^{-5}$ .

#### 4. Conclusion

We presented a measure for estimating the amount of multilingual noise in monolingual corpora. It can be calculated efficiently as it involves only 1000 frequency counts per noise language tested. Experiments show that the measure correlates well with artificial mixtures of monolingual corpora. For large corpora, noise will be detected down to a ratio of  $10^{-5}$ .

A possible application in a World Wide Web context is to measure the amount of web sites that belong to a defined set of languages. This is done by querying the index of a search engine for the top 1000 words per language for frequency to produce statistics as e.g. in (Langer 2001).

#### 5. References

- Biemann, Chr., Bordag, S., Heyer, G., Quasthoff, U. and Wolff, Chr. (2004): *Language-independent Methods for Compiling Monolingual Lexical Data*. Proceedings of CicLING 2004, Seoul, Korea and Springer LNCS 2945, pp. 215-228, Springer Verlag Berlin Heidelberg
- Brown, P.F., Della Pietra, V. J., deSouza, P., Lai, J.C. and Mercer, R. L. (1992): *Class-Based n-gram Models of Natural Language*. Computational Linguistics 18(4):467-479
- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990): *Indexing by latent semantic analysis*. Journal of the Society for Information Science, 41(6):391-407
- Dunning, T. (1994): *Statistical Identification of Language*. Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University
- Goldsmith, J. (2001): *Unsupervised learning of the morphology of a natural language*. Computational Linguistics, 27:153-198
- Kilgarriff, A. (2001): *Web as corpus*. In Proceedings of Corpus Linguistics 2001, Lancaster, England.
- Langer, S. (2001): *Natural languages on the World Wide Web*. In: Bulag. Revue annuelle. Presses Universitaires Franc-Comtoises, S. 89-100
- Leech, G. (1992): *100 million words of English: the British National Corpus*. Language Research 28:1, 1-13.
- Quasthoff, U., Biemann, C. and Richter, M. (2006): *Corpus Portal for Search in 16 Monolingual Corpora*. Proceedings of LREC-2006, Genoa, Italy
- Resnik, P. and Smith, N.A. (2003): *The Web as a Parallel Corpus*. Computational Linguistics 29(3):349-380
- Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Addison-Wesley, Reading, MA.