Disentangling from Babylonian Confusion – Unsupervised Language Identification

Chris Biemann, Sven Teresniak

Leipzig University
Computer Science Institute, NLP Dept.
Augustusplatz 10/11
04109 Leipzig, Germany
biem@informatik.uni-leipzig.de, knorke@zehnvierzig.org

Abstract: This work presents an unsupervised solution to language identification. The method sorts multilingual text corpora on the basis of sentences into the different languages that are contained and makes no assumptions on the number or size of the monolingual fractions. Evaluation on 7-lingual corpora and bilingual corpora show that the quality of classification is comparable to supervised approaches and works almost error-free from 100 sentences per language on.

1 Introduction

With a growing need for text corpora of various languages in mind, we address the question of how to build monolingual corpora from multilingual sources without providing training data for the different languages.

According to [Pantel et al. 2004], shallow methods of text processing can yield comparable results to deep methods when allowing them to operate on large corpora. The larger the corpus, however, the more difficult it is to ensure sufficient quality. Most approaches in Computational Linguistics work on monolingual resources and will be disturbed or even fail if a considerable amount of 'dirt' (sublanguages or different languages) are contained. Viewing the Internet as the world's largest text corpus, it is difficult to extract monolingual parts of it, even when restricting downloading to country domains or some domain servers.

While some languages can be identified easily due to their unique coding ranges in ASCII or UNICODE (like Greek, Thai, Korean, Japanese and Chinese), the main difficulty arises in the discrimination of languages that use the same coding and some common words, as most of the European languages do.

In the past, a variety of tools have been developed to classify text with respect to its language. The most popular system, the *TextCat Language Guesser* as described in [Cavnar & Trenkle 1994], makes use of the language-specific letter N-gram distribution and can determine 69 different natural languages. According to [Dunning 1994], letter trigrams can identify the language almost error-freely from a text-length of 500 bytes on. Other language identification approaches use short words and common words as features, e.g. [Johnson 1993], or combine both approaches (cf. [Schulze 2000]). For a comparison, see [Grefenstette 1995].

All these approaches work in a supervised way: given a sample of each language, the model parameters are estimated and texts are classified according to their similarity to the training texts. But supervised training has a major drawback: The language identifier will fail on languages that are not contained in its training and, even worse, it will mostly have no clue about that and assign some arbitrary language[1].

This work proposes a method that operates on words as features and finds the number of languages as well as the sentences that belong to each language in a fully unsupervised way. Of course, the method is not able to tell the names of the involved languages, but rather groups sentences of similar languages together.

2 Methodological Approach

In this section we describe our methodology. With the use of co-occurrence statistics we construct weighted lexeme graphs built from words as nodes and their associations as edges. A graph algorithm determines sub-graphs with high connectivity (clusters) and assigns class labels to each word. Under the assumption that two words of the same language exhibit more frequent co-occurrence than word pairs of different languages, the graph algorithm will find one cluster for each language. The words in the clusters serve as features to identify the languages of the text collection by using a sentence-based language identifier.

2.1 Sentence-based Co-occurrence Graphs

The joint occurrence of two or more words within a well-defined unit of information (sentence, document or word window) is called a co-occurrence. For the selection of significant co-occurrences, an adequate measure has to be defined: Our significance measure is based on the Poisson distribution: Given two words *A*, *B*, each occurring *a*, *b* times in sentences, and *k* times together, we calculate the significance *sig(A, B)* of their occurrence in a sentence as follows:

$$sig(A,B) = \frac{x - k\log x + \log k!}{\log n}$$

$$\text{with } n = \text{number of sentences in the corpus,}$$

$$x = \frac{ab}{n}$$

Roughly speaking, this significance measure is the negative logarithm of the probability for seeing at least k joint occurrences if A and B would be statistically independent.

[1] e.g. TextCat (*http://odur.let.rug.nl/~vannoord/TextCat/Demo/*) assigns “Nepali” to texts like “xx xxx x xxx …” and “Persian” to “öö ö öö ööö …”

If the significance value is above a certain threshold (we use 0.4), the co-occurrence of *A* and *B* is considered significant. How significant co-occurrences can serve as a data basis for a variety of applications is described in [Biemann et al. 2004a] and [Biemann et al. 2004b].

The entirety of all words having significant co-occurrences can be viewed as nodes in a graph. An edge between two words exists, if their co-occurrence is significant in the text, and the weight of the edge is given by the significance value.

Figure 1 shows the co-occurrence graph calculated from the text of this paper from section 1 until the significance measure formula. For texts that short, significant co-occurrences do not reflect semantic relations as they do in large corpora (see [Quasthoff & Wolff 2002]).

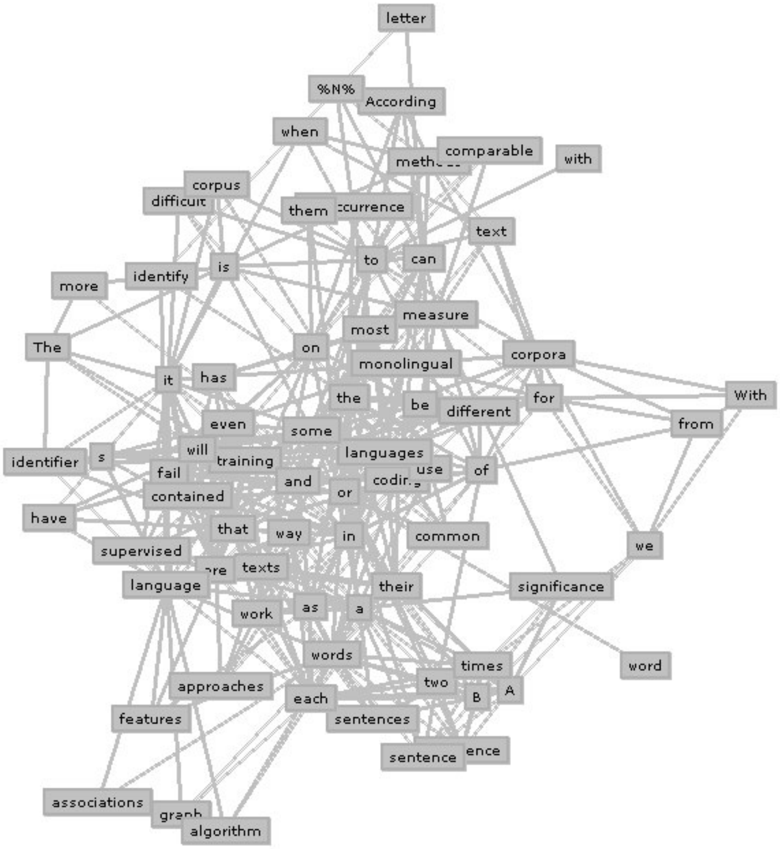


Figure 1: Co-occurrence graph for the beginning of this document. The visualization software positions associated words close to each other. Note the cluster containing *two, times, A* and *B*. Edge weight is not reflected in the visualization.

The structure of co-occurrence graphs can be characterized by the small-world property: a high clustering coefficient paired with short path lengths between nodes and a number of scale-free properties (cf. [Ferrer-i-Cancho & Sole 2001], [Barabási et al. 2000].) As a consequence, there exist sub-graphs that are almost completely connected (clusters), and some hubs (nodes with a high connection degree) that connect those clusters with each other.

When calculating co-occurrence graphs for multilingual corpora, one cluster for each language can be expected (that has in itself again the small world property), while frequent words belonging to different languages serve as hubs.

2.2 Finding Clusters in Co-occurrence Graphs

In the following we describe a graph algorithm called Chinese Whispers[2], which finds clusters in (co-occurrence) graphs by assigning class labels for each cluster. The name is motivated by the children's game where a message is passed by whispering amongst several children, transforming the message into something funny. In this case, the nodes of the graph whisper their label to each other, until every node agrees with its adjacent nodes on some label.

Figure 2 gives an outline of the algorithm:

```
Assign different labels to every node in the graph;

For iteration i from 1 to total_iterations {

          mutation_rate= 1/ (i^2);

          For each word w in the graph {

            label of w = highest ranked label in
                                 neighbourhood of w;

            with probability mutation_rate:
                  label of w = new class label;

          }

}
```

Figure 2: Outline of the Chinese Whispers algorithm, which finds monolingual clusters in co-occurrence graphs of multilingual corpora

The algorithm begins by assigning different class labels to all words in the graph. In the main loop, the class labels are subject to change: Every word inherits the highest ranked class label in its neighbourhood, which consists of all nodes that are connected to the word in question. Ranking is done using the weights of the edges

[2] Thanks goes to Vincenzo Moscati for the name of the algorithm.

that are given by the significance value of the co-occurrence: For each class label in the neighbourhood, the sum of the weights of the edges to the word in question is taken as score for ranking. Figure 3 illustrates the change of a class label. Note that all changes are not immediately applied, but take effect in the next iteration.

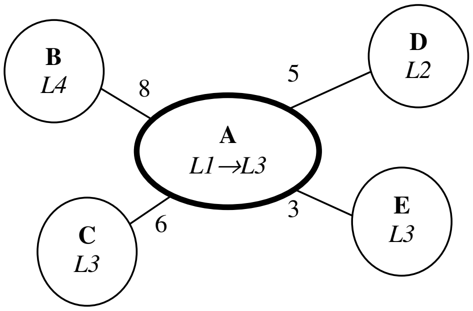


Figure 3: The class label of word A changes from *L1* to *L3* due to the following scores in the neighbourhood: *L3*:9, *L4*:8 and *L2*:5.

With increasing iterations, clusters become self-preserving: If a strongly connected cluster happens to be homogenious with respect to class labels, it will never be infected by a few connections from other clusters.

To avoid premature convergence and scenarios, where several clusters end up with the same class labels due to strong influence of hubs, a mutation step is included that assigns new class labels with decreasing probability dependent on the iteration. In principle, the algorithm works without mutation on large graphs as well but its performance decreases on small graphs. Although random influence is introduced, only a very small part of the nodes are labeled differently in different runs on the same graph.

The number of total iterations was set to 20 for all experiments, because after 20 iterations the changes in the class labels were only minimal, even in graphs of 1 million nodes and more.

2.3 Sentence-based Language Identification

For sentence-based language identification we use a tool called “LanI“ („Language Identifier“) which we developed with the aim to cleanse the sentences and corpora downloaded from the WWW from alien material like source code, sublanguages, foreign languages and so on as described above. LanI was written in Python. Here sentence-based means that LanI returns acceptable results on strings containing at least four or five words. Thereby only statistical data are used, so no semantic or syntactic information is utilized and the well-formedness of sentences is not essential. By designing LanI we assumed, that the probability for a sentence to belong to a given language is computable with the knowledge of the relative frequency of the words of the given sentence in different languages. Considering Zipf's Law [Zipf

1929], LanI gets by with the 250 to 10,000 high-frequency words of given languages. So a word that only appears in one language *L* (accordingly in one language-wordlist), of course increases the probability for a given sentence to be written in that language *L*. Needless to say, word lists are not disjoint, for example in borrowings like the English "kindergarten" and German "Kindergarten" or like "[to] die" in English and "die" (an article) in German. If so, the relative frequency for in both languages matters. For example: the relative frequency for "die" in English is approx. 0.0000367 (0.004% of all words in written English) but around 0.029 (2.9%) in German. Thus, a sentence with "die" contributes more for German (and for English only a little bit).

For the relative frequency for words in different languages we used corpora collected in "Projekt Deutscher Wortschatz", *http://wortschatz.uni-leipzig.de/*). The relative frequencies are regarded as word probabilities $P_L(w)$ for every word *w* in language *L*.

After normalisation and smoothing (to avoid multiplication with zero, every word not occurring in a wordlist but in another, gets a very small fixed value) we compute the sentence-probability by multiplying the relative probability of each word. For a sentence *S* we compute for every language *L*:

$$P_L(S) = P_L(w_1) \cdot P_L(w_2) \cdot \ldots \cdot P_L(w_s)$$

The result of this is, that we have a probability for every language and now we can easily select the best language. Does the winning language have a likelihood (at least) as twice as large than the second best language, the language for the given sentence is determined. If only ten percent or less of the words of a sentence contained in the wordlists, the calculation will be discarded and LanI reports "unknown".

However, for the algorithm described in this paper we modify the approach a little: for every word in each wordlist the probability is set to a given constant value, thus there is no attempt to use distribution information. For each cluster we got, the word lists consist of the items of each of it. Because every word in the graph is assigned to at most one cluster by the graph algorithm, there are no words contributing to the assignment of two or more languages.

3 Experiments and Evaluation

In order to evaluate our approach, we applied the method to two different settings. A corpus compiled from equal fractions of seven European languages served to determine the lower threshold of how many sentences of each language should be at least contained to make the method work. To find out, how much the amount of monolingual material can differ in biased corpora, experiments on bilingual corpora with different size factors between the two contained languages were performed.

We used the standard Information Retrieval measures for evaluation: *Precision* (P), which is the number of true positives divided by the sum of true positives and false positives, and *recall* (R), which is calculated by dividing the number of true positives through the number of total target items. For some experiments, we provide the harmonic mean of P and R, the *F-value* (F), which is given by (2PR)/(P+R).

In all experiments, clusters were assumed to represent languages if they contained at least 1.8% of all words in the co-occurrence graph. All smaller clusters were regarded as noise. This value was determined empirically by observations on monolingual corpora.

3.1 Multilingual Test Corpora

From a sample of 100'000 sentences each of the languages Dutch, Estonian, English, French, German, Icelandic and Italian we compiled multilingual corpora that contained 100, 200, 500, 1'000, 5'000, 10'000, 50'000 and 100'000 sentences of each language and performed the calculation of co-occurrences. Smaller versions are completely contained in larger versions, average sentence length is 127 characters. Figure 4 illustrates the number of nodes and edges in the co-occurrence graphs for each 7-lingual corpus.

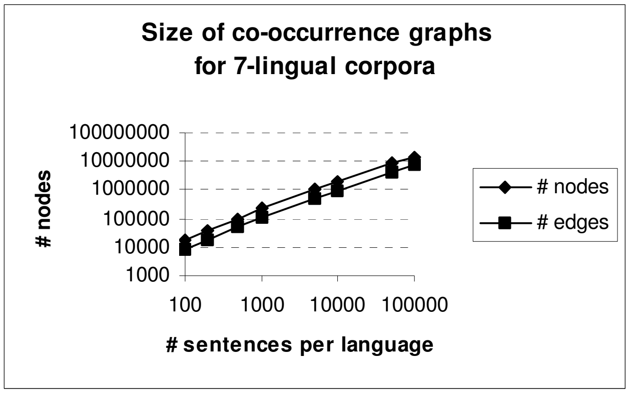
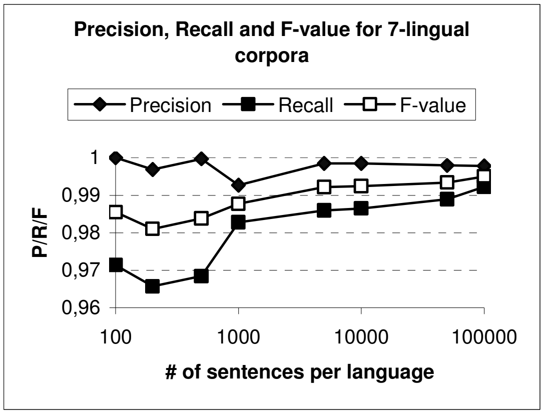


Figure 4: The sizes of co-occurrence graphs dependent on corpus size. Both axes in logarithmic scale.

After applying the graph algorithm, all experiments resulted in seven large clusters of words that ended up with the same class label, as well as many very small clusters that never exceeded 1.7% of the total number of nodes in the graph. Figure A1 and A2 in the appendix illustrate the effect of the graph algorithm on the co-occurrence graph of the smallest 7-lingual corpus.

Precision and Recall varied slightly for the different languages. Whereas English was the easiest language to identify (Precision > 99.9%, Recall > 98.1% for all experiments), Estonian was the most difficult (Precision > 99.3% for all experiments, Recall 86.7% for 100 sentences, 92.7% for 200 sentences and > 93.7 for all other experiments). All other languages were classified close to the quality of English.

Figure 5 presents the overall results from the 7-lingual corpora experiments.



Size	Prec	Rec	Fval
100	1	0.9714	0.9855
200	0.9969	0.9657	0.9810
500	0.9997	0.9684	0.9838
1K	0.9927	0.9828	0.9877
5K	0.9985	0.9859	0.9922
10K	0.9985	0.9865	0.9924
50K	0.9979	0.9889	0.9922
100K	0.9978	0.9922	0.9950

Figure 5: Results for 7-lingual corpora of different sizes.

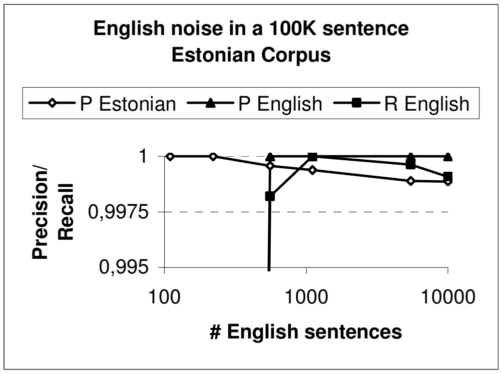
For error analysis, we looked at the reasons for low recall in small corpora and checked misclassified items. The main recall flaw was caused by the Estonian parts, which is rooted in the nature of the Estonian corpus. Whereas the other monolingual fractions are taken from newswire sources, our Estonian originates from legal texts. A fraction of about 3% consists of 'sentences' indicating dates or law paragraph ciphers, like "*10.12.96 jõust.01.01.97 - RT I 1996 , 89 , 1590.*" Other unclassified sentences in all languages were mostly enumerations of sport teams or short headlines.

The few misclassified sentences mainly contained proper names, in many cases company names that were usually classified as English. In some cases, the language identifier picked the wrong language for bilingual sentences, like French for "*Frönsku orðin "cinéma vérité" þýða "kvikmyndasannleikur".*" in the Icelandic section.

Experiments with fewer than 100 sentences for each of the seven languages resulted in more than seven (mostly about eleven) clusters that could not group the sentences of one language together via the language identifier. The significant co-occurrences were too small in numbers and too noisy in quality.

3.2 Bilingual Corpora

While the experiments in the previous section focussed on equal fractions of many languages, we now describe how the method behaves on bilingual corpora with monolingual fractions of differing sizes. This setting is somewhat more realistic when identifying languages in web data, for a domain usually provides most of its content in one language and translates only a few parts. In order to examine the minimum size of 'dirt' language our method can notice, we performed experiments with biased mixtures of Estonian and English, German and Dutch, Italian and French. The first language contributed the main part comprising of 100'000 sentences, the amount of the second language was varied from 100, 200, 500, 1'000, and in the English-Estonian case 5'000 and 10'000 sentences. It turns out that a second language is identified if it contributes 500 or more sentences. Figure 6 depicts the results for the English-Estonian mixture.

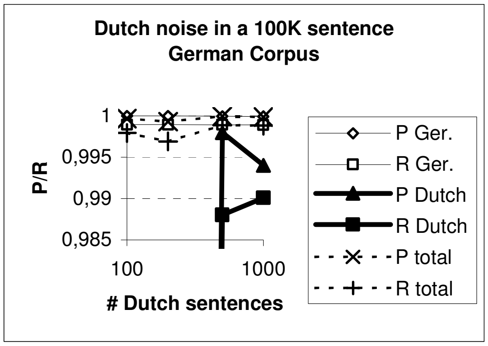


Eng. sent.	Eng P	Eng R	Est P	Est R
100	0	0	1	0.9678
200	0	0	1	0.9674
500	1	0.9982	0.9996	0.9664
1K	1	1	0.9994	0.9674
5K	1	0.9996	0.9989	0.9674
10K	1	0.9991	0.9989	0.9674

Figure 6: Results for different noise levels of English in a 100’000-sentence Estonian corpus. Estonian recall is not depicted for scaling reasons.

The low recall on Estonian was caused by the same reasons as mentioned in the previous section. As we expected, languages from as different families as English and Estonian are very well separable, at least if the size factor is below 200. For less than 500 English sentences, the cluster for English became too small to distinguish it from other small Estonian clusters. Nevertheless, 80% of the English sentences were classified as unknown.

More difficulties were to be expected dealing with language of the same family, like the two Germanic languages German and Dutch. Figure 7 presents the results:

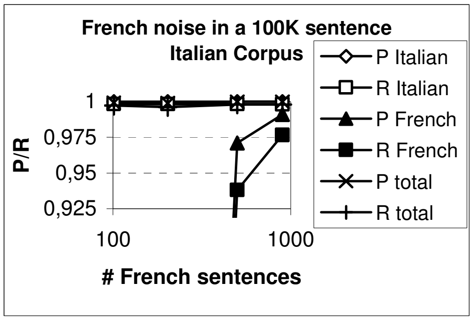


Dut.. sent.	100	200	500	1000
Dut. P	0	0	0.9980	0.9940
Dut. R	0	0	0.9880	0.9901
Ger. P	1	1	0.9999	0.9999
Ger. R	0.9990	0.9990	0.9989	0.9989
Total P	0.9997	0.9993	0.9999	0.9999
Total R	0.9979	0.9969	0.9989	0.9988

Figure 7: Results for Dutch noise in a 100’000-sentence German corpus.

In the two experiments where Dutch could not be identified, 66% of the Dutch sentences were regarded unknown, one third was classified as German.

Two languages that have even more common words were examined in the same setting: the Romanic languages Italian and French, see figure 8. Results differ slightly from the other language pairs: already at 500 sentences of French, figures are decreasing. The similarity of these two languages is also reflected in that about 60% of French sentences were considered to be Italian in the two small-fraction experiments.



Fre. sent.	100	200	500	1000
Fre. P	0	0	0.9710	0.9910
Fre. R	0	0	0.9380	0.9767
Ita. P	1	1	1	0.9999
Ita. R	0.9983	0.9983	0.9983	0.9982
Total P	0.9972	0.9988	0.9999	0.9999
Total R	0.9973	0.9962	0.9980	0.9980

Figure 8: Results for French noise in a 100'000-sentence Italian corpus.

Experiments show that similar languages are only slightly more difficult to separate than languages of different families, as long as the bias towards the main language in the corpus is not larger than 200.

In further experiments, no language pairs were found that could not be separated under the conditions described. Even German dialects could be identified in large German corpora.

4 Conclusion and Further Work

The main contribution of our work is that we show that unsupervised language identification is possible and works with the same high accuracy as the well-known supervised approaches mentioned in the introduction. The only requirement on the language data is the possibility to recognise word boundaries (which even might be replaced by using chunks of fixed length for e.g. Chinese) and sentences (which also might be replaced by using chunks of e.g. 20 words) for the calculation of significant co-occurrences. The method is robust with respect to the number and the mass distribution of the involved languages and produces reliable results from 100 sentences per language on.

When sorting document collections by splitting the documents into sentences, applying our approach and assigning the majority of sentence-languages to the document, there should be virtually no errors.

In further work, we will examine how the method performs on other document classification tasks, like identification of web genres (cf. [Rehm 2002]) or text classification on standard resources like [Reuters 2000]. For this, we will have to modify the graph algorithm in order to find small clusters that reflect genre-specific wording.

References

[Barabási et al. 2000] Barabási, A.L., Albert, R., Jeong, H. (2000): Scale-free characteristics of random networks: the topology of the World-wide web, Physica A (281)70-77

[Biemann et al. 2004a] Biemann, C., Bordag, S., Heyer, G., Quasthoff, U., 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

[Biemann et al. 2004b] Biemann, C., Böhm, K., Heyer, G., Melz, R. (2004): Automatically Building Concept Structures and Displaying Concept Trails for the Use in Brainstorming Sessions and Content Management Systems, Proceedings of I2CS, Guadalajara, Mexico

[Cavnar & Trenkle 1994] Cavnar, W. B., J. M. Trenkle (1994): N-Gram-Based Text Categorization. In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175

[Dunning 94] Dunning, T. (1994): Statistical Identification of Language. in Technical report CRL MCCS-94-273, Computing Research Lab, New Mexico State University, March 1994.

[Ferrer-i-Cancho & Sole 2001] Ferrer-i-Cancho, R. and Sole, R. V. (2001) The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1482):2261--2265

[Grefenstette 1995] Grefenstette, G. (1995): Comparing Two Language Identification Schemes. The proceedings of 3rd International Conference on Statistical Analysis of Textual Data (JADT 95), Rome, Italy, Dec. 1995.

[Johnson 1993] Johnson, S. (1993): Solving the problem of language recognition. Technical Report, School of Computer Studies, University of Leeds

[Quasthoff & Wolff 2002] Quasthoff, U.; Wolff, C. (2002): The Poisson Collocation Measure and its Applications. In: Proc. Second International Workshop on Computational Approaches to Collocations, Wien.

[Pantel et al. 2004] Pantel, P., Ravichandran, D., Hovy, E. (2004) : Towards Terascale Semantic Acquisition, Proceedings of the 20th International Conference on Computational Linguistics (COLING-04), Geneva, Switzerland

[Rehm 2002] Rehm, G. (2002): Towards Automatic Web Genre Identification. Proceedings of the 35th Hawaii International Conference on System Sciences, Hawaii.

[Reuters 2000] Reuters Corpus (2000). Volume 1, English language, http://about.reuters.com/researchandstandards/corpus

[Schulze 2000] Schulze, B.M. (2000): Automatic language identification using both N-gram and word information. US Patent No. 6,167,369.

[Zipf 1929] Zipf, G. K. (1929): Relative Frequency as a Determinant of Phonetic Change, Reprinted in Harvard Studies in Classical Philology, Volume XI.

Appendix: Figures

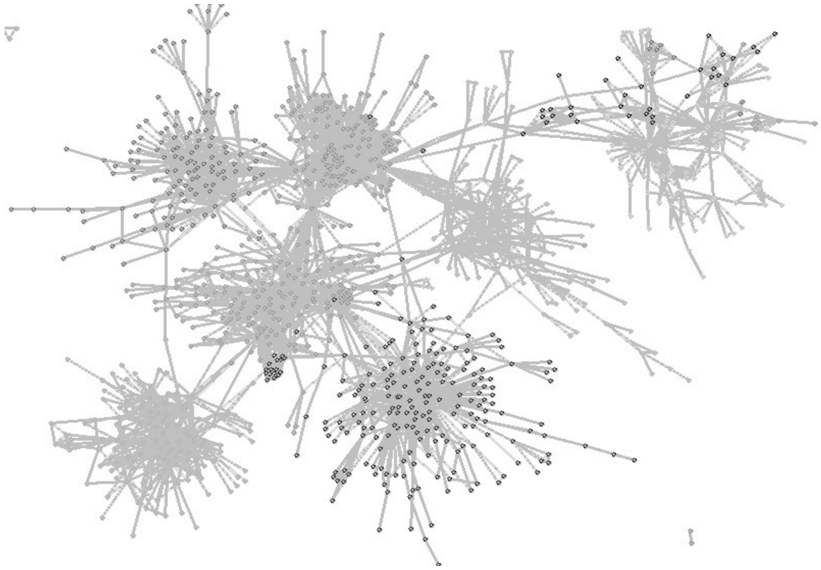


Figure A1: The co-occurrence graph of the 7-lingual corpus with 100 sentences per language, coloured according to the class labels. Black nodes in the right upper corner do not belong to one of the seven large clusters obtained by Chinese Whispers.

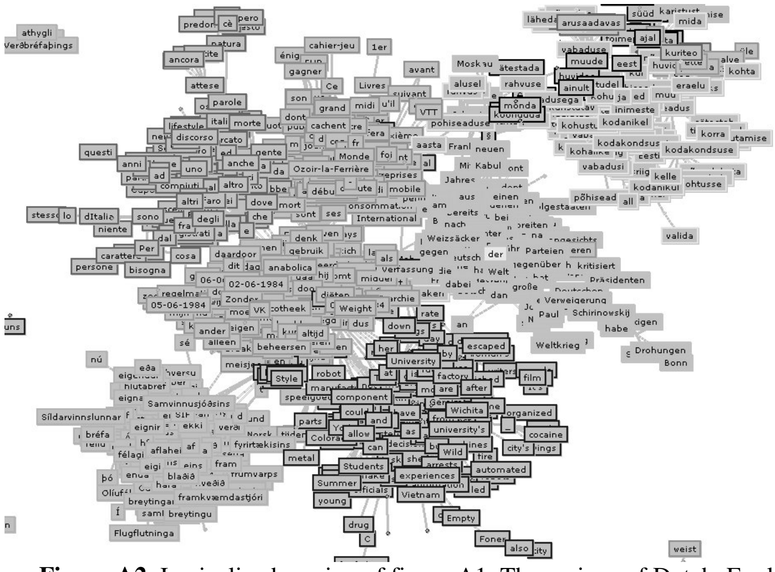


Figure A2: Lexicalized version of figure A1. The regions of Dutch, English, Estonian, French, German, Icelandic and Italian are clearly visible.