Vocabulary-Based Language Similarity using Web Corpora

Dirk Goldhahn, Uwe Quasthoff
Natural Language Processing Group, University of Leipzig, Germany
E-mail: dgoldhahn@informatik.uni-leipzig.de, quasthoff@informatik.uni-leipzig.de

Abstract

This paper will focus on automatic methods for quantifying language similarity. This is achieved by ascribing language similarity to the similarity of text corpora. This corpus similarity will first be determined by the resemblance of the vocabulary of languages. Thereto words or parts of them such as letter n-grams are examined. Extensions like transliteration of the text data will ensure the independence of the methods from text characteristics such as the writing system used. Further analyzes will show to what extent knowledge about the distribution of words in parallel text can be used in the context of language similarity.

Keywords: language similarity, corpora, vocabulary

1. Introduction

The detection of similar languages is a concern of different scientific disciplines. Language typology is interested in general similarity or in a more restricted kind according to certain properties. Fields like language identification are interested in languages which are similar based on their vocabulary in order to identify difficult pairs of languages which are confusable.

In this paper we investigate different fully automatic approaches to language similarity, which allow us to quantify the likeness of languages using just their vocabulary. So we aim at automatic measurements of a distance between languages.

Basis of our studies will be Web corpora of the Leipzig Corpora Collection (LCC) (Goldhahn, 2012). Since only few text samples are available for many languages, robustness of our approach when using small corpora is necessary. Expandability when adding new languages is also an issue.

By assigning the similarity of languages to the similarity of corpora and their vocabulary we are able to compute distances for every language pair we can obtain text for. The only further requirement is that we are able to tokenize textual data for languages in question.

At first we have a closer look at similarity on an orthographic level. Techniques used in fields such as language identification will be utilized here. To that end features such as the most frequent words or parts of words like letter n-grams will be examined and compared. Different similarity measures and weightings of the features will be evaluated. Influence of textual properties like subject area will be examined. Especially the described evaluations are a novelty in this context.

The approach will be extended by transliteration to allow for comparison beyond script boundaries.

Results are compared to genealogical language relationships to receive an objective evaluation of similarity. Since we aim at also finding language pairs which are similar despite not being in a genealogical relationship, this is not the optimal data to evaluate against. But yet, algorithms capable of identifying unknown pairs of similar languages should also be able to reproduce this kind of relation. In addition results of the analyses will also be evaluated by hand.

Furthermore parallel text corpora are utilized for language comparison. By analyzing cross-language vocabulary distribution, we will determine language similarity. In addition, by searching related words on orthographic and phonologic level among similarly distributed pairs, we enhance this approach further.

2. Related work

Since Greenberg's work (1963) in typology languages are mainly categorized according to certain structural features. These properties have to be determined manually and are not always known for a high percentage of the world's languages. Classifications based on typological features complement divisions of languages based on genealogical or spatial relatedness.

Other typological studies are concerned with language comparison based on manually created word lists. Lists of a base vocabulary covering 100 concepts (Swadesh, 1952, 1955) are translated into many languages and form a starting point for orthographic or phonetic pair wise comparison which leads to statements about language similarity. Glottochronology extends this research further by manually (Swadesh, 1950, 1955, 1971) or automatically (see Embleton, 1986 for an overview; Brown, 2008) determining an approximate date when related languages diverged from each other.

Other studies evaluate measures of a distance between words (Wichmann, 2010), but normally only within one script. Kondrak and Sherif (2006) determine cognates among word pairs by computing phonetic similarity.

Automatic language identification is a task closely related to language similarity. It is concerned with assigning a text to the closest known language by comparing features such as common words or n-grams (Cavnar, 1994; Dunning, 1994; Grefenstette, 1995). In contrast to language identification, language similarity aims at general statements about languages and not about single texts. Hence, influences of textual properties like subject area are treated in this paper. Furthermore it is not only identifying the closest language but considers relations to all languages. Eventually it can be helpful to identify

problematic language pairs that can be easily confused. Language identification already has techniques for identifying such pairs of languages such as confusion matrices. However, such approaches require a large amount of documents for each language. Word list based methods manage to do so with very little text.
There are few works using parallel text for language comparison. Mayer (2012) uses word alignment and matrix algebra for this task.

3. Word list based approaches

3.1 Data

Web corpora of the LCC for 346 languages are used in the following experiments. Text collections from different sources are compiled to test robustness of the approaches in regard to factors such as subject area or text type. We utilize the Universal Declaration of Human Rights (UDHR) containing about 1800 words (English version). Further textual data are added by utilizing Watchtower texts. In addition random web corpora are generated for well resourced languages.

3.2 Methods

In this section languages are compared based on lists of common vocabulary. Different automatic processing steps are necessary to achieve this.
First, we extract profiles from text corpora, which are used for later comparison. These profiles consist of:

- the most frequent words or
- the most frequent letter trigrams.

We utilize lists of different length in our experiments.

In a next step a similarity value has to be calculated for each pair of languages based on common vocabulary in the sense of identical strings. There is a wide choice of possible measures used in text clustering (Huang, 2007). They are applied to the profiles described before.
We utilize rank correlation coefficients or vector space based techniques such as:

- Kendall tau distance for ranked lists (Kendall, 1938) with an own extension for lists with unequal sets of elements
- Cosine similarity and
- Dice coefficient, the number of common elements, as a baseline.

Dependent on the similarity measure used, different weightings of the elements of the extracted profiles are possible. Rank correlation uses the rank of elements while Dice coefficient does not rank them at all. When using cosine similarity different weightings are possible, we use:

- (reversed) rank,
- frequency and
- logarithm of the frequency.

The latter is applied due to the typical distribution of word frequencies in natural languages. The use of the logarithm diminishes the influence of very high frequencies in the top words according to Zipf's Law.
To evaluate our results we compare them to the genealogical classification of languages on levels such as family or genus. For this purpose we cluster the resulting similarity matrix of all languages using just the number of occurring genealogical classes as further input. We then determine cluster purity of our solution as a measure of correctness of our solution. Since we do not solely aim at rebuilding language families, we also have a manual look some results.
Figure 1 depicts all the steps necessary for computing language similarity.

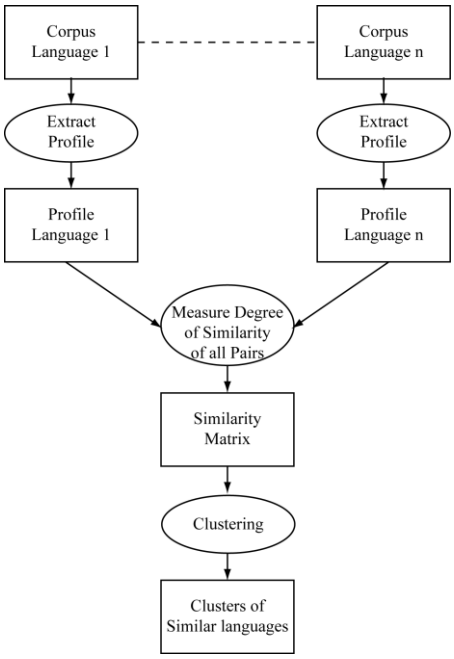

Figure 1: Stages of computation of language similarity.

3.3 Results

Results of the analyses can be found in figure 2. Obviously language comparison based on simple orthographic profiles can lead to results which resemble e.g. genealogical relations between languages. Apparently trigrams yield better results compared to words. Since frequent trigrams often correspond to typical word constituents such as prefixes or affixes this is reasonable. Even between closely related languages there might be only few common words, but many mutual trigrams.

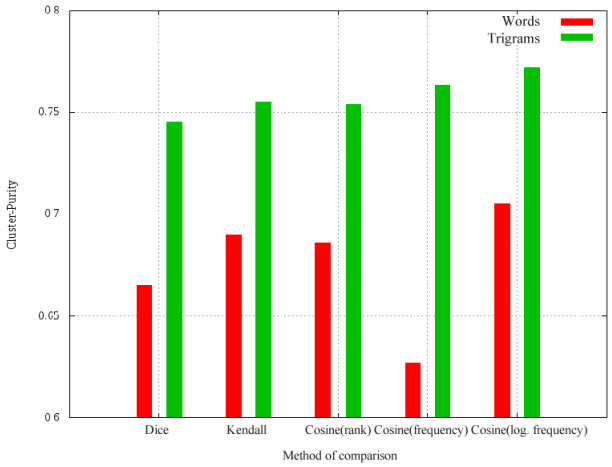

Figure 2: Cluster-Purity for word- and trigram-based language comparison dependent on weighting of the features and measure of similarity. Values are computed with reference to 108 language genera present in the data.

Influences of different methods of comparison and feature weighting can also be seen. Use of logarithmic frequencies is beneficial independent of the profile applied. Utilizing pure frequencies has negative effects when comparing word lists.

The effect of methods and weighting used is also visible when looking at examples as in figures 3 and 4.

Rank-based comparison leads to a correct solution containing only North Germanic languages. The use of the Dice coefficient leads to a cluster erroneously including English and an English-based Pidgin language as can be seen in figure 4. This is based on many common words with same spelling but different meaning. Such words are often alike by chance. In table 1 the most frequent words of the Swedish corpus, which also appear in the corpus of Nigerian Pidgin, can be found. Among them are many word pairs with different meanings which have identical word forms by coincident. One example is the word *far*. In Nigerian Pidgin it has the same meaning as in English but in Swedish it means 'father'.

Rank-based approaches help to overcome these problems by penalizing large rank differences as they occur in table 1. Table 2 depicts common words of corpora of Swedish and Icelandic. As opposed to the previous example most identical words have a common meaning in both languages and rank differences are typically low. Thus rank-based techniques will identify these languages to be very similar.

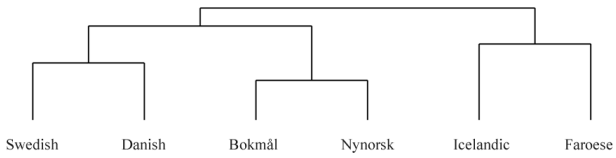

Figure 3: Clustering solution of North Germanic languages based on cosine similarity of word ranks.

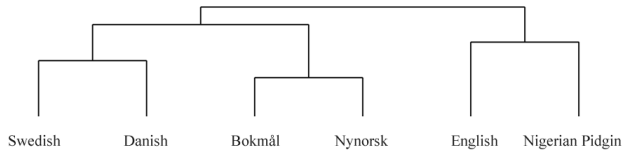

Figure 4: Clustering solution of North Germanic languages based on Dice coefficient.

Word rank in Swedish corpus	***Word rank in Nigerian Pidgin***	***Word***
23	157	man
39	1	dem
123	197	in
325	43	person
383	324	god
386	14	be
436	22	of
454	24	all
463	201	form
598	98	far
623	318	information
805	55	bad
831	12	and
836	383	december
888	36	ting

Table 1: The most frequent common words of two corpora in Swedish and Nigerian Pidgin.

Word rank in Swedish corpus	Word rank in Icelandic corpus	Word
6	23	en
21	66	sig
24	27	var
72	151	alla
78	67	upp
100	33	hans
104	113	vill
120	144	allt
140	57	fram
181	277	kom
194	211	annan
247	500	enda

Table 2: The most frequent common words of two corpora in Swedish and Icelandic.

So far we were only able to detect similar languages within the same writing script. By enhancing our approach with transliteration we can overcome these boundaries. For that we integrated components from ICU[1]. ICU allows for language-independent transliteration based just on the script used. Thus, extension of analyses to new languages is simple, as long as no new script is involved. In return overall transliteration quality is lower. Figure 5 shows a cluster of our solution containing just Slavic languages. When using transliteration resulting clusters successfully bypass script boundaries.

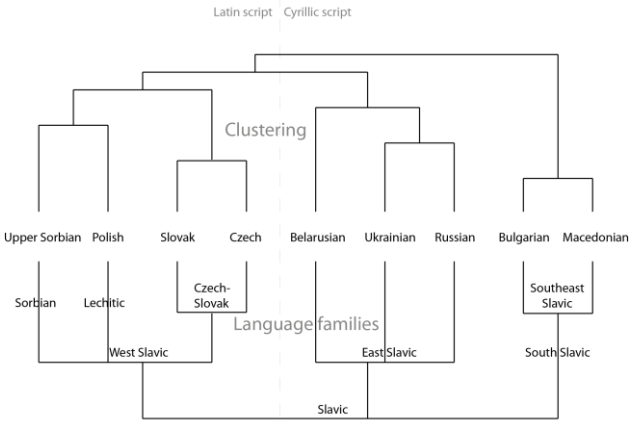

Figure 5: Clustering solution of Slavic languages in Latin script in comparison to genealogical relations. Results are based on cosine similarity of transliterated trigrams weighted by rank.

Figure 6 shows the influence of the length of lists used for comparison. Small lists seem to suffice. Once again results are dependent on the weighting of the features.

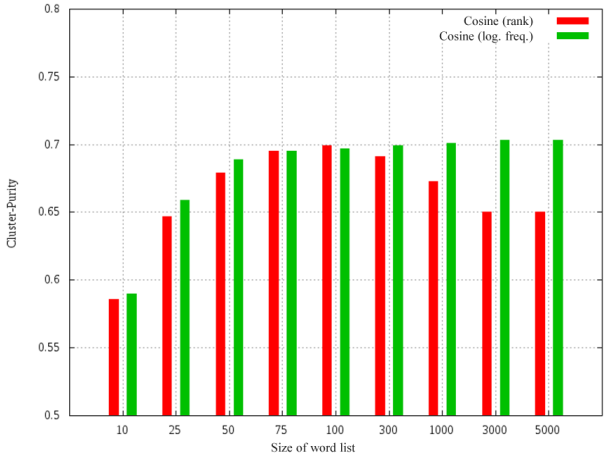

Figure 6: Cluster-Purity for word-based language comparison dependent on size of the word list and weighting of the features.

The dependency of our results on general characteristics of the texts can be seen in figure 7. Word-based comparison yields worse cluster purity when corpora from different subject areas are used. Trigrams seem to be less susceptible to such textual properties.

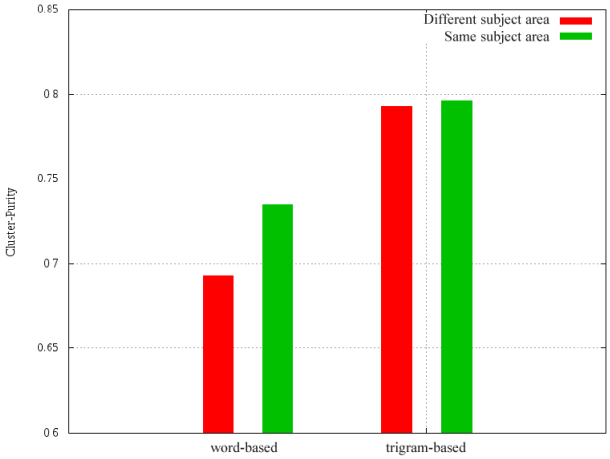

Figure 7: Cluster-Purity for word-based and trigram-based language comparison dependent on subject areas of the underlying corpora.

[1] International Components for Unicode: http://www.icu-project.org

4. Approaches based on parallel text

4.1 Data

There are different sources for parallel text, among them UDHR or religious texts like Watchtower or the Bible[2]. Due to the availability in many languages and the extent of about 8,000 verses in the New Testament, Bible texts are used in the following studies.

4.2 Methods

Using parallel text such as Bibles it is possible to align corresponding words across languages (Melamed, 1996; Biemann, 2005). This is solely based on the cross-language distribution of words across sentences. This results in a value of accordance for every cross-language word pair.

On the basis of a genealogically stratified sample of 160 languages these values are created for all words of all language pairs and form the starting point for further analyses.

In a first investigation we use the amount of word pairs above a certain threshold to estimate language similarity. This is based on the assumption that similar languages should have similarly distributed words.

In a second approach we examine these word pairs in more detail. We identify cognates (words with similar form because of common etymological origin) among them using orthographic or phonetic comparison. For that we apply Levenshtein distance and phonologically weighted Levenshtein distance with naive transliteration into the International Phonetic Alphabet (Kondrak, 2000, 2006). The amount of cognates is then utilized as a measure of similarity.

4.3 Results

Figure 8 depicts the results of analyzing the count of translation pairs. Once again our results resemble genealogical relations quite well. But some results vary strongly compared to word list based approaches. Using lists of words or trigrams there are no languages similar to Korean. Based on translation pairs we find Altaic languages to be close to Korean (table 3). This is supported by Miller (1971) and Song (2005) who identify grammatical similarities between them, indicating our approach is discovering morphological or syntactical similarities between languages.

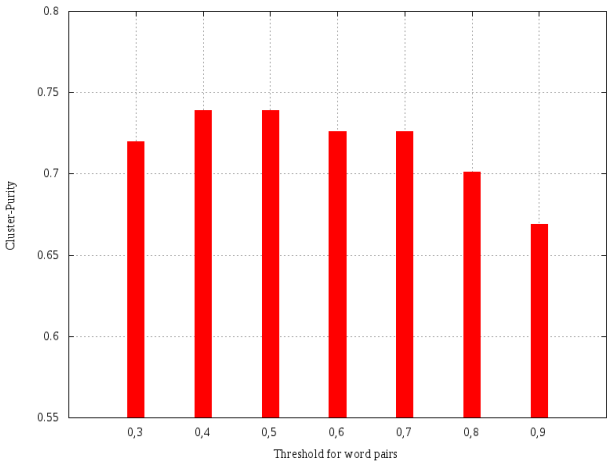

Figure 8: Cluster-Purity for language comparison based on number of translation pairs. Results are depicted dependent on a threshold for word similarity.

Rank	*Language 1*	*Language 2*	*Family*	*Script*
1	Korean	Kazakh	Altaic	Kyrillic
2	Korean	Uzbek	Altaic	Kyrillic
3	Korean	Turkish	Altaic	Latin

Table 3: Languages most similar to Korean when using language comparison based on translation pairs.

When comparing the cognate-based approach to previous approaches like the word list based ones, we observe a higher quality of results. These analyses reproduce genealogical relations better than other methods. At that phonetic comparison is slightly ahead of the orthographic one (figure 9).

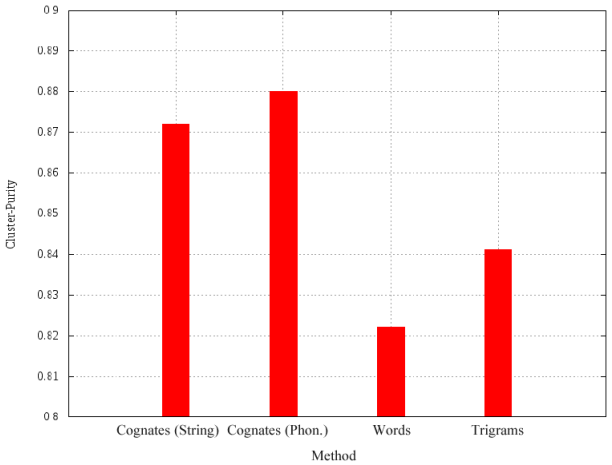

Figure 9: Cluster-Purity for different approaches of language similarity. On the left cognates among translational equivalents are analyzed, while on the right word-based and trigram-based language comparison (as in section 3) is performed.

[2] http://www.bible.is

5. Conclusion

This paper successfully introduced and evaluated vocabulary-based approaches for determining language similarity. All methods were automated and based solely on text corpora, whereby the proposed procedure differs from other work such as Swadesh-based approaches.

First we utilized methods working with lists of frequent words or letter trigrams to compare languages on an orthographic level. We identified trigrams to be more suitable for this task when matching results with genealogical relationships. Thus, typical constituents of words seem to be a better indicator for similarity. Furthermore, trigrams proved to be more stable concerning textual properties like subject area, making them the more versatile feature.

In addition different measures and feature weightings were compared. Rank- and frequency-based approaches turned out to be more robust than unweighted ones. However, principles like zipf's law have to be taken into consideration.

In the second part of the paper parallel text was utilized for language comparison. Analyzing just the distribution of words, an approach independent of the orthographical form of words was introduced. We were able to show that this method captures some kind of grammatically-based similarity between languages setting it apart from other techniques.

Finally cognates were identified among similarly distributed words on a phonetic or orthographical level. In comparison to genealogical relations this approach proved to be the best performing, thus highlighting parallel text as an excellent source for determining language similarity.

6. References

Biemann, C., Quasthoff, U. (2005): Dictionary acquisition using parallel text and cooccurrence statistics. Proceedings of NODALIDA 2005,Joensuu, Finland.

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world′ s languages: a description of the method and preliminary results. STUF-Language Typology and Universals, 61(4), 285-308.

Embleton, S. (1986). Statistics in Historical Linguistics [Quantitative linguistics, vol. 30]. Bochum: Brockmeyer.

Cavnar, W.; Trenkle, J. (1994). N-Gram-Based Text Categorization. Proceedings of the Third Annual Conference on Document Analysis and Information Retrieval (SDAIR), Las Vegas, pp. 161-175.

Dunning, T. (1994). Statistical identification of language. Technical Report MCCS 94–273, New Mexico State University.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In Proceedings of the Eighth Language Resources and Evaluation Conference (759-765).

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 2, 73-113.

Grefenstette, G. (1995). Comparing two Language Identification Schemes. JADT 1995, 3rd International conference on Statistical Analysis of Textual Data, Rome.

Huang, A. (2007), Similarity Measures for Text Document Clustering. In Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC'08), Christchurch, New Zealand.

Kendall, M. (1938). A New Measure of Rank Correlation. Biometrika 30 (1–2): 81–89.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In Proceedings of NAACL 2000, pages 288–295.

Kondrak, G.; Sherif, T. (2006). Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. Proceedings of the COLING-ACL Workshop on Linguistic Distances, 43-50, Sydney, Australia, July 2006.

Mayer, T., & Cysouw, M. (2012). Language comparison through sparse multilingual word alignment. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH (pp. 54-62). Association for Computational Linguistics.

Melamed, I. D. (1996). Automatic construction of clean broad-coverage translation lexicons. In Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas, Montreal, Canada.

Miller, R. A. (1971). Japanese and the other Altaic languages (p. 222). Chicago: University of Chicago Press.

Song, J. J. (2005). The Korean language: Structure, use and context. Routledge.

Swadesh, M. (1950). Salish internal relationships. International Journal of American Linguistics, 16(4), 157-167.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. Proceedings of the American philosophical society, 96(4), 452-463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. International journal of American linguistics, 21(2), 121-137.

Swadesh, M. (1971). The Origin and Diversification of Language. Chicago/New York: Aldine-Atherton.

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. Physica A: Statistical Mechanics and its Applications, 389(17), 3632-3639.