Corpus-Based Linguistic Typology: A Comprehensive Approach

Dirk Goldhahn **Uwe Quasthoff** **Gerhard Heyer**
Natural Language Processing Group,
University of Leipzig,
Germany
dgoldhahn,quasthoff,heyer
@informatik.uni-leipzig.de

Abstract

This paper will have a holistic view at the field of corpus-based linguistic typology and present an overview of current advances at Leipzig University. Our goal is to use automatically created text data for a large variety of languages for quantitative typological investigations. In our approaches we utilize text corpora created for several hundred languages for cross-language quantitative studies using mathematically well-founded methods (Cysouw, 2005). These analyses include the measurement of textual characteristics. Basic requirements for the use of these parameters are also discussed. The measured values are then utilized for typological studies. Using quantitative methods, correlations of measured properties of corpora among themselves or with classical typological parameters are detected. Our work can be considered as an automatic and language-independent process chain, thus allowing extensive investigations of the various languages of the world.

1 Introduction

Text corpora are a versatile basis for linguistic analyses. They allow for investigations of different aspects of languages, among others grammatical levels like morphology or syntax. The World Wide Web is a comprehensive source used for the creation of corpora. One advantage of using Web text is the availability of data for a large variety of languages. Since linguistic typology is concerned with cross-linguistic universals of language, Web corpora are an attractive source for investigations in this field. But as of today, few typological studies make use of automatically created text corpora or even Web based corpora.

Our goal is to use the textual data of the Leipzig Corpora Collection (LCC) (Quasthoff, 1998; Quasthoff, 2006; Biemann, 2007) in typological studies. Hence, we created an automatic and language independent process chain, which includes all steps necessary for this intention. As a first step this involves the acquisition and creation of the text corpora of LCC for several hundred languages. Following this, various measurements of different complexity such as average word length or average number of syllables per word are taken on these corpora. In addition constraints for their application for typological analyses are determined. Finally these measurements are utilized for corpus-based typological analyses applying quantitative methods. We determine correlations between measurements and dependencies between measurements and typological parameters like morphological type or position of case marking. This also allows to predict such classical typological features such as word order. Focus of the paper will be on general methodologies. Simple examples will be presented to illustrate the possibilities of the approaches. They will elucidate that even when using a fully automatic analysis, which in many cases will be very superficial, general characteristics of languages can still be captured. By allowing for broad analyses using large datasets this approach can complement existing typological methodologies and form a basis for further manual inspection and interpretation.

2 Acquisition and creation of Web corpora

Web corpora of LCC form the basis of the following typological investigations. LCC offers access to corpus-based monolingual full form dictionaries via a Web interface, Web services and allows for the download of text data. Corpora for more than 200 languages are available online[1]. The dictionaries contain statistical information for each word of the corpus.

All in all, corpora for more than 1,000 languages have been created which will be used for the following analyses. Corpus sizes vary from several hundred million to about 8000 sentences (languages where only the New Testament exists). Because of copyright issues many of them cannot be made available online. This holds, among others, for Bible texts. All corpora and their corresponding statistics are created from web pages. Thus, a process chain for the automatic acquisition, creation and statistical evaluation of corpora from web sources has been implemented which is presented in Goldhahn (2012).

3 Corpus-based statistics

3.1 Measurements

This paper aims at using simple corpus-based statistics for typological studies. Thus, measurements on the corpora have to be taken. Therefore, all in all more than 100 features are measured for each corpus using an automated framework. The measurements are conducted on different levels of language like sentences, words or letters. These levels are easy to identify for nearly all languages. Thus they are ideal for comparable studies. Other measurements are concerned with entities, which are more difficult to determine, such as syllables.

Among the measurements are features such as:

- Average word length in characters
- Average sentences length in words or characters
- Text coverage of the most frequent words
- Different measurements of vocabulary richness
- Entropy on word and character level
- Average number of syllables per word or sentence
- Average syllable length
- Ratio of prefixes and suffixes

For a more complete list of possible features see Goldhahn (2013).

Some values can be determined quite easily like the average sentence length in words, as long as word and sentence boundaries are identified correctly. Most languages use white space to separate words which allows for easy segmentation. Few languages are lacking clear word boundaries in their written form (e.g. Chinese and Japanese). In such cases specific tools such as Stanford Word Segmenter[2] are used.

For few measurements, only a rough approximation is possible. Examples are features concerned with prefixes and suffixes. Without analyzing the morphological processes of a language in detail, assertions about affixes are difficult. Therefore we chose to consider typical word beginnings and endings. Among them we identified those which discriminate many otherwise identical word pairs. This appeared to be a good approximation for affixation in many languages. Syllables also turned out to be difficult entities to measure, mainly because of the varying use of certain letters as vowels or consonants as well as the use of diphthongs depending on language. Therefore we used the language independent algorithm of Sukhotin (1988) to determine vowels and consonants of each language using text samples. The number of syllables of a word is equal to the number of syllable peaks. For most languages the number of peaks is close to the number of vowels in a word, since diphthongs are normally rarely encountered. On this basis statistics concerned with syllables can be computed und used for further analysis.

3.2 Constraints on measurements

Measurements on text corpora depend on different factors such as:

- Language

[1] http://corpora.uni-leipzig.de

[2] http://nlp.stanford.edu/software/segmenter.shtml

- Preprocessing
- Measurement process
- General characteristics of the texts like genre, text type or text size

In this paper we are mainly interested in changes of certain measured properties dependent on language. The preprocessing and the process of measurement, which can also have an influence on the resulting values (Eckart, 2012), are kept identical at all time. But other general characteristics of the corpora differ greatly between the texts in question. Therefore not every measured value is useful for typological analyses. Certain constraints have to be met or considered to enable proper insights from typological experiments.

First of all we examined the relative standard deviation (SD) of the measurements between languages. Especially in case of roughly approximated measurements a high cross-language variance can improve results of statistical tests used in the following typological analyses. Table 1 depicts relative SD for different measurements.

In addition we inspected the influence of textual characteristics such as text type, subject area or corpus size. This can lead to limitations concerning the usability of different corpora for certain measured parameters. Hence, it is desirable to have a higher cross-language SD in comparison to these other textual properties. Examples for these comparisons can be found in Table 2.

Furthermore, classical typological parameters normally vary less within groups of genealogically related languages. Since we aim at relating our measurements to typological features, the same is expected for our measurements. Table 3 shows, that this is generally the case. Exceptions in certain language families - just as word length in this case - can be subject for further investigations.

Negative results in these analyses do not necessarily exclude certain statistics from further investigations. But one has to assume that they will have impact on the results of typological studies conducted. As an example, the well-known Type-Token-Ratio (TTR), which measures the ratio of the number of different words forms to the number of total words in a text, is examined. It is well known that TTR is susceptible to changes of text size. When conducting a study with corpora of varying size, this will probably reduce the statistical significance of the results or even produce invalid results. One solution is to unify the amounts of text of the different corpora used. Since this results in throwing away valuable data, it is often an alternative to modify the statistic in question. Type-Token-Ratio is a measure of vocabulary richness. However, other measures of this property such as Turing's Repeat Rate, which measures the average number of words until a random word in the text occurs again, are hardly influenced by corpus size.

4 Typological analyses

4.1 Linguistic typology

Linguistic typology is concerned with the classification of languages according to their structural properties. This allows for the identification of possible and preferred structures of language. On the one hand typology determines typological parameters used for language classification. On the other hand it examines regularities or universals, which these parameters follow. Among them are relationships between different typological features (Greenberg, 1963).

4.2 Methods

In this section we relate simple features of text corpora, which can be determined using automatic means, with classical typological parameters of language, which describe different levels of language such as morphology or syntax. Furthermore we try to relate different measured features. We use quantitative methods like correlation analysis (Pearson product-moment correlation coefficient) and tests of significance (Wilcoxon, 1945; Mann, 1947) to analyze and confirm such relationships (Cysouw, 2005). In addition we predict typological parameters using methods of supervised machine learning or Bootstrapping approaches. In comparison to other works in this field (Fenk-Oczlon, 1999) the process does not contain any manual steps. We use automatically generated text resources combined with an automatic measurement process. Together they allow for the analysis of big textual data in several hundred languages while

Measurement	Average	Relative SD
Average word length (Types)	9.11	0.37
Average word length (Tokens)	5.55	0.46
Average sentence length in words	26.87	0.27
Average sentence length in char.	161.98	0.23
Ratio of suffixes and prefixes	4.10	1.96
Text coverage of top 100 words	56.82	0.21

Table 1: Average values and relative standard deviation for corpus-based measurements.

Measurement	$\frac{\text{SD(Language)}}{\text{SD(Corpus size)}}$	$\frac{\text{SD(Language)}}{\text{SD(Text type)}}$	$\frac{\text{SD(Language)}}{\text{SD(Subject area)}}$
Average sentence length in words	107.41	8.65	13.20
Average sentence length in characters	77.032	6.23	7.67
Ratio of suffixes and prefixes	18.78	17.69	25.84
Syllables per sentence	30.25	8.22	7.33
Type-Token-Ratio	1.16	8.21	6.13
Turing's Repeat Rate	238.95	6.37	8.69
Text coverage of the top 100 words	530.85	7.93	8.75

Table 2: Comparison of standard deviations of corpus-based measurements. Values larger than 1 imply a higher cross-language standard deviation compared to the standard deviation when varying other features such as corpus size. Values much larger than 1 are desirable.

Measurement	$\frac{\text{SD(Random)}}{\text{SD(Germanic)}}$	$\frac{\text{SD(Random)}}{\text{SD(Indo-European)}}$	$\frac{\text{SD(Indo-European)}}{\text{SD(Germanic)}}$
Average word length (types)	5.22	0.71	7.38
Average word length (tokens)	4.51	0.72	6.26
Average sentence length in words	5.16	2.30	2.24
Average sentence length in characters	3.61	2.37	1.52
Type-Token-Ratio	2.54	1.80	1.41
Turing's Repeat Rate	6.46	2.09	3.08
Ratio of suffixes and prefixes	4.70	2.71	1.73
Text coverage of the top 100 words	4.18	1.33	3.14

Table 3: Comparison of cross-language standard deviations between language groups of different coherence. A random sample of languages is compared to a sample of Indo-European or Germanic languages. In general values larger than 1 are expected and imply a higher standard deviation in the less coherent language group.

considering a high number of features and possible relations.

4.3 Results

4.3.1 Correlations between measurements

We were able to detect several correlations between measured parameters of corpora. By applying correlation analysis to comparable corpora in 730 languages we achieved results of high statistical significance and found interesting correlations or confirmed known ones. Some of them will be presented in this section. Since the focus of this paper is on methodology, only very brief interpretations of results will be offered. Such results can be a starting point for typologists that need to analyze each language in detail in order to accomplish a full interpretation.

We found:

- A negative correlation between average length of words and average length of sentences (in words): $Kor_e = -0.55, p < 0,001\%$, sample size of 730.

The longer the average word of a language is, the fewer words are usually utilized (or needed) to express a sentence.

- A negative correlation between average number of syllables per word and average number of words per sentence: $Kor_e = -0,49, p < 0,001\%$, sample size of 730. The more syllables the average word of a language has, the fewer words are typically used to express a sentence.

- A negative correlation between text coverage of the most frequent words and entropy on word level: $Kor_e = -0.97, p < 0,001\%$, sample size of 730. When few words are very frequent, the average expected value of the information contained in a single word usually becomes lower.

4.3.2 Relationships between measurements and typological parameters

We also determined various relations between measured parameters and classical typological parameters using tests of significance. Typological information was taken from the World Atlas of Language Structures (WALS[3]) (Cysouw, 2007b). Since typological data is sparse, sample sizes are usually smaller than 730 languages.

Only a small sample of all results which were achieved is presented in this paper. For a full overview see (Goldhahn, 2013).

We found a significant relation between ratio of suffixes and prefixes and position of case marking (end of word vs. beginning of word):

- $p < 0.001\%$, mean values of 10.48 and 0.7 and sample sizes of 57 and 11.

Our simple automated measurements regarding affixes are sufficient to capture a relation to actual processes of affixation in languages. Although we obviously measure more than just case marking a significant relation can still be established. It seems that case marking has a big influence on our measurement.

Morphological type

We found:

- A significant relation between average length of words of a language and its morphological type (concatenative vs. isolating): $p < 1\%$, mean values of 8.43 and 6.95 and sample sizes of 68 and 8.

- A significant relation between measured amount of affixation of a language and its morphological type (concatenative vs. isolating): $p < 0.5\%$, mean values of 21.20 and 10.06 and sample sizes of 68 and 8.

- A found a significant relation between entropy on word level of a language and its morphological type (concatenative vs. isolating): $p < 0.05\%$, mean values of 9.95 and 8.64 and sample sizes of 68 and 8.

Several measurements we conducted, among them those concerning average word length or affixation, are related to morphological features of languages. Opposing features such as concatenative and isolating morphological type are presented as an obvious example.

Syllables

We confirmed results of Frank-Oczlon (1999) for a considerably larger sample of languages with higher significance. We also enriched these results with further findings. Among others we discovered significant relations between:

- Average number of syllables per sentence and word order (SOV vs. SVO): $p < 0.001\%$, mean values of 56.95 and 45.27 and sample sizes of 111 and 137.

- Average number of syllables per word and morphological type (concatenative vs. isolating): $p < 5\%$, mean values of 2.06 and 1.64 and sample sizes of 68 and 8.

4.3.3 Prediction of typological parameters

Typological parameters have been determined for many languages of the world

[3] http://wals.info/

and can be looked up in collections such as WALS. But for many parameters only partial knowledge is available. Hence, ways to predict typological features based on automatic measurements would be of great help.

Supervised machine learning

One way to predict typological parameters is the use of supervised machine learning. To illustrate the possibilities of this approach the example of morphological type of a language will be discussed. Once again only concatenative and isolating languages will be analyzed, which form two extremes concerning morphological properties. Table 4 shows the probabilities of correct classification of morphological type using 76 languages which WALS assigned to one of these two classes. As input different measured features or combinations of them were utilized. Especially when using a mix of several features for prediction, high accuracies of over 90% were achieved.

Applying this method the usability of a high number of corpus features for predicting different typological parameters can be investigated.

Features used	**Correctly classified**
Words per sentence	74.40%
Number of word forms	87.76%
Words per sentence, number of word forms, syllables per word	91.84%

Table 4: Probability of correct prediction of morphological type (concatenative vs. isolating) using a Support Vector Machine based on different feature sets.

Bootstrapping

Furthermore we utilized automatic Bootstrapping approaches to determine typological parameters. Some information such as part of speech of words (POS), which is necessary to predict certain typological parameters, can only be assigned automatically for few well-resourced languages. Using parallel text such as Bibles it is possible to align corresponding words across languages (Melamed, 1996; Biemann, 2005; Cysouw, 2007a). This knowledge can then be used to spread information about features like POS to further languages. By applying graph partitioning algorithms such as Chinese Whispers (Biemann, 2006) we were able to successfully transfer this information about POS to languages without known POS-Tagger. This knowledge together with information about translational equivalents was then used to predict the typological parameter of word order. Therefore word order information of a source language (German) was transferred to the target languages using sample sentences (see Figure 1). This way we were able to successfully determine the correct word order for sample languages. See Goldhahn (2013) for details about the methodologies of this example, such as the use of a simplified Tagset.

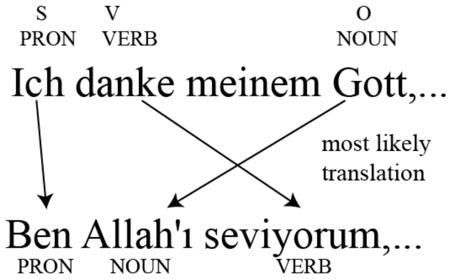

Figure 1: Depiction of the information used to predict word order in a target language.

5 Conclusion

In this paper, we presented a novel approach to corpus-based linguistic typology allowing for a new kind of typological analyses. Using an automatic process chain we were able to measure statistical features of corpora of web text for several hundred languages. These properties were applied in quantitative typological analyses to detect correlations with classical typological parameters or to predict such parameters. Several simple results were presented. They give insight into the possibilities of the methodologies described in this paper and show that despite using a superficial automatic approach general characteristics of languages can still be captured. By adding further features that can be measured automatically or by analyzing relationships to additional typological parameters a wide area of typological issues can be investigated. Since this approach facilitates broad analyses of very large datasets it can complement existing typological work and form a basis for further manual inspection and interpretation.

References

Baroni, M.; Bernardini, S. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004.*

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

Biemann, C. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing* (pp. 73-80). Association for Computational Linguistics.

Biemann, C.; Heyer, G.; Quasthoff, U.; Richter, M. 2007. The Leipzig Corpora Collection - Monolingual corpora of standard size. *Proceedings of Corpus Linguistic 2007*, Birmingham, UK.

Cysouw, M. 2005. Quantitative methods in typology. In Altmann, G.; Khler, R.; Piotrowski, R. (eds.). *Quantitative linguistics: an international handbook*, 554 - 578. Berlin: Mouton de Gruyter.

Cysouw, M.; Wlchli, B. 2007a. Parallel texts: using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung*, 60(2), 95-99.

Cysouw, M. 2007b. Special issue on analyzing the World Atlas of Language Structures. *Sprachtypologie und Universalienforschung*.

Eckart, T.; Quasthoff, U.; Goldhahn, D. 2012. The Influence of Corpus Quality on Statistical Measurements on Language Resources. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012).*

Fenk-Oczlon, G.; Fenk, A. 1999. Cognition, Quantitative Linguistics, and Systemic Typology. *Linguistic Typology*, 3: 151 - 177.

Goldhahn, D.; Eckart, T.; Quasthoff, U. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012).*

Goldhahn, D. 2013. Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken. Dissertation, University of Leipzig, Leipzig, Germany.

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

Mann, H. B.; Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1).

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

Quasthoff, U. 1998. Projekt der deutsche Wortschatz. Heyer, G.; Wolff, Ch. (eds.), *Linguistik und neue Medien*, Wiesbaden, pp. 93-99.

Quasthoff, U.; Richter, M.; Biemann, C. 2006. Corpus Portal for Search in Monolingual Corpora. *Proceedings of LREC 2006.*

Sukhotin, B. V. 1988. Optimization algorithms of deciphering as the elements of a linguistic theory. *Proceedings of the 12th conference on Computational linguistics-Volume 2* (pp. 645-648). Association for Computational Linguistics.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6), 80-83.