Statistical Corpus and Language Comparison using Comparable Corpora

Thomas Eckart, Uwe Quasthoff
NLP Group, University of Leipzig
Johannisgasse 26, 04103 Leipzig, Germany
E-mail: {teckart, quasthoff}@informatik.uni-leipzig.de

Abstract

Corpora of different languages but similar genre allow language comparison. Applying the same methods to corpora of the same language but of different genre or origin results in corpus comparison. Having many corpora in identical formats, these statistical methods will generate various data for manual or automatic analysis. The introduced system reports more than 150 results per corpus, for approximately 150 corpora right now. The results are presented on more than 22,000 pages which are generated automatically. Intelligent Browsing allows contrasting of different corpora with respect to different questions, languages, text genres and varying corpus size. As a side effect, shortcomings in the corpus preprocessing usually produce statistical anomalies that are easily noticeable and lead to an improved processing chain.

1. The Leipzig Corpora Collection

Basis for all further considerations are the corpora of the Leipzig Corpora Collection. For about fifteen years corpora are created by using text material of all kind, focusing on the Internet as text resource. By using the Web text material in more than 50 languages and in partially enormous sizes were gathered from various sources.

By now hundreds of corpora were created, which can be classified in three dimensions: language (including dialects), genre (currently: news texts, random web texts, governmental and Wikipedia texts) and size (measured in number of sentences). For easy corpus comparisons, subcorpora of normed sizes (containing 10,000, 30,000, …, 3 million sentences), are created.

All texts are segmented into sentences and words and all relevant data is stored in a relational database (cf. Quasthoff et al., 2006), containing information like word frequencies and word co-occurrences. To ensure comparability, the corpus preprocessing was standardized as much as possible (cf. Quasthoff & Eckart, 2009). Currently, corpora in 15 languages are made freely available, an extensive expansion of the download portal is planned for the near future[1].

2. Analysis Procedure

With a standardized creation process and a uniform data schema on the one hand and a fast growing amount of different corpora on the other, it became obvious that there was a lack of analysis tools to evaluate existing data and to ensure corpus quality without extensive manual work. As a result, existing tools (mostly Python and Perl scripts of different complexity) were replaced by a new tool with the intention to separate the knowledge- and labor intensive creation of an evaluation task from the execution of this task on a specific corpus.

Therefore every evaluation is encapsulated in a single script, that holds all necessary information and that validates against a proprietary XML schema. In general, one script consists of a set of SQL statements that are executed on a database, specified by the user. Each result set can be processed further by the scripting languages Perl or PHP, including: merging of data, reformatting of result sets or computing interesting values that couldn't be provided by the database management system itself. These data are sufficient for many problems of corpora analysis. To offer more intuitive ways, especially in the field of statistical evaluation, a graphical component is needed. Hence, the plotting tool Gnuplot[2] was integrated, that offers various possibilities of graphical presentation.

To ensure platform independence only software was used that is provided for different platforms and systems, namely Java, PHP, Perl and Gnuplot. Additionally an

[1] *http://corpora.informatik.uni-leipzig.de/download.html*

[2] *http://www.gnuplot.info*

Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, pages 15–20
Malta, 22 May 2010

easy-to-use Graphical User Interface was developed, and the possibility of executing a set of evaluation scripts in a batch mode.

3. Analysis Types

To cover as many fields of interest as possible, more than 150 different evaluation scripts were created and classified in six sections of analysis:

Corpus Meta Information

Information regarding the corpus and its creation: size, versions of preprocessing tools, duration of the processing tool chain etc.

Characters and Character N-Grams

Information regarding the distribution of characters, especially on word beginnings or endings, character successor rates, character transition probabilities etc.

Words and Multi-words

Information regarding words (including multi-words if existing): length distribution, text coverage, samples, several variants of Zipf's law (cf. Zipf, 1949), word transition probabilities, word similarity using Levenshtein distance, average word length, longest words in different frequency ranges etc.

Sentences

Information regarding distribution of sentence lengths measured in words or characters, typical sentence beginnings or endings, similar sentences, sentences containing only words of either high or low frequency etc.

Word Co-occurrences

Samples for typical word (sentence / neighbour-) co-occurrences (cf. Dunning, 1994), visualization of Zipf's law for co-occurrences, semantic word similarity using joint co-occurrences, small world parameters for the co-occurrence graph etc.

Sources

Information regarding sources like: number of used sources, typical size of each source, differences between various sources measured in parameters as above, etc.

These fields are steadily extended and will be developed further. The focus here is especially on customization and extension of existing scripts to character sets and syntactic structures that haven't been dealt with yet.

4. Language and Corpora Comparison

4.1 General Structure

An analysis script as described above usually generates three different types of output:

- A table containing the measured data, together with a Gnuplot diagram
- One or two parameters (like the slope for Zipf's law) to approximate the function plotted above
- Example corpus data for extreme data points (for Zipf's law: the most frequent words)

These three distinct output types can be used for different purposes: a plotted diagram is fine for manual inspection and manual corpora comparison. Numeric parameters are more interesting for automatic comparisons: the parameters of different analysis can be considered as components of a feature vector for a corpus. Clustering techniques can then be used to identify families of similar corpora or languages.

Sample words or sentences with extreme parameters are of interest due to their specific linguistic properties or may help to find corpus preprocessing problems, as will be shown below.

4.2 Intra-language and Inter-language Comparisons

While language dependent parameters are expected to vary for different languages, their behavior for different genres within one language is difficult to predict. The following table compares three parameters first for different text genres of German, and then the same parameters for newspaper corpora for different languages. The intra-language variation may help to decide whether differences between languages can be considered as significant. Moreover, for corpora of mixed or unknown genre such data help to decide whether more detailed information about the genres are necessary.

	Text coverage (20 top words)	**Avg. word length**	**Avg. sentence length**
News	22.10%	13.59	16.19
Web	21.57%	14.06	16.03
Wikipedia	23.10%	12.57	16.71
Movie Subtitles	21.20%	10.42	6.57

Table 1: Intra-language comparison

Table 1 and 2 show the text coverage for the 20 most frequent words, the average word length in characters (without multiplicity) and the average sentence length in words.

	Text coverage (20 top words)	Avg. word length	Avg. sentence length
German	22.10%	13.59	16.19
English	26.23%	10.62	19.46
Czech	16.78%	8.65	14.95
Vietnamese	12.44%	4.97	23.64
Finnish	12.37%	12.28	11.50

Table 2: Inter-language comparison

4.3 Insights into Language Structure

The following example counts the number of letter n-grams as a measure for character successor variability. Because rare words (especially when containing spelling errors) will contain nearly any n-gram, only the N=10^k most frequent words (for k=2, 3, 4, …) are used.

Table 3 shows the number of different letter n-grams at word beginnings, taken from a newspaper corpus in Finnish.

N	# of bigrams	# of 3-grams	# of 4-grams	# of 5-grams
100	51	82	95	99
1000	211	449	654	822
10000	577	1821	3256	4829
100000	1391	6852	16804	28622
1000000	2512	14910	44494	86492

Table 3: Finnish n-grams at word beginnings

In figure 1, the values of table 3 are plotted with logarithmic scale. The nearly straight lines suggest a power law.
Similar results are true for counting letter n-grams at word endings or counting letter n-grams regardless of their position. The same is true for many other languages. Of course, the slope varies for the different n-gram types and languages.

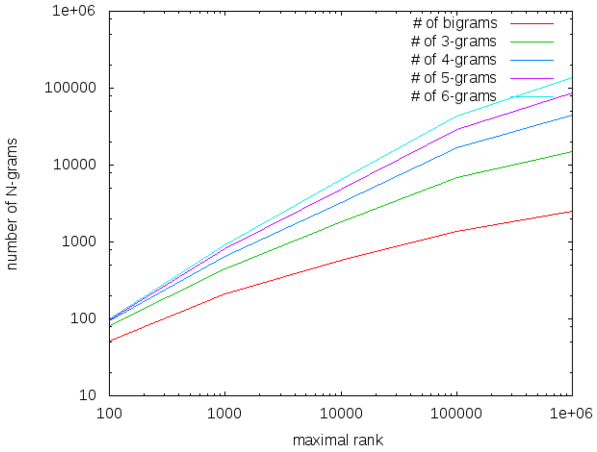


Figure 1: Letter n-grams of Finnish word beginnings

4.4 Non-linear Growth Rates

The non-linear growth of certain parameters gives rise to new difficulties when comparing different corpora or languages. For such comparisons we can use the corpora of normed size as explained in section 1. Figure 2 shows the number of distinct word forms, the number of sentence based word co-occurrences and the number of next neighbor co-occurrences. These numbers are taken for corpora of 100.000, 300.000, 1 million and 3 million sentences. Again, the nearly straight lines imply a power law. A more detailed inspection using different languages still shows nearly straight lines, but with slightly different parameters.

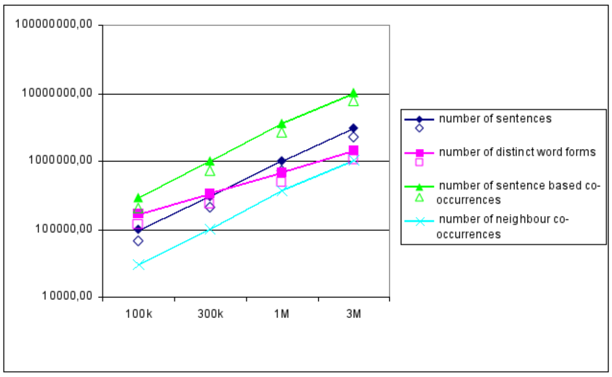


Figure 2: Non-linear growth

5. Quality Assurance

In many of the above mentioned analysis types, a special value is measured for many objects, like sentence length in characters for every sentence of a corpus. Looking at objects with extreme values (i.e. very small or very large), we often find effects of errors in the input material or poor preprocessing (cf. Eskin, 2000; Dickinson & Meurers, 2005).

In the case of very short sentences, we may find broken sentences. Moreover, sentences containing many very low frequent words are usually not well-formed. Table 4 shows sentences of an English Web corpus that consists of words that have a very low average frequency. Apparently there were encoding problems in the input material and the language identification failed in rejecting some non-English sentences.

Avg. word rank	**Sentence**
35844	Bidh an luchd-aithisg a' gabhail notaichean tron choinneimh agus 's dÃ²cha […] briathran air an togail.
31711	"HCPT - The Piligrimage Trust" jest organizacja charytatywna zalozona w Wielkiej Brytanii.
28524	Gjelder dette for barn og unge mennesker under 18 Ã¥r?

Table 4: Examples of sentences that consist of words with low average frequency

Another hint for problems in the corpus generation process is looking at extreme points of the distribution of specific characters.

# of semicolons	**# of sentences**
2	183
3	28
4	16
5	4
6	1
12	1

Table 5: Part of a semicolon distribution in Ukrainian sentences

Table 5 shows an excerpt of the distribution of semicolons in a 100,000 sentences Ukrainian corpus.

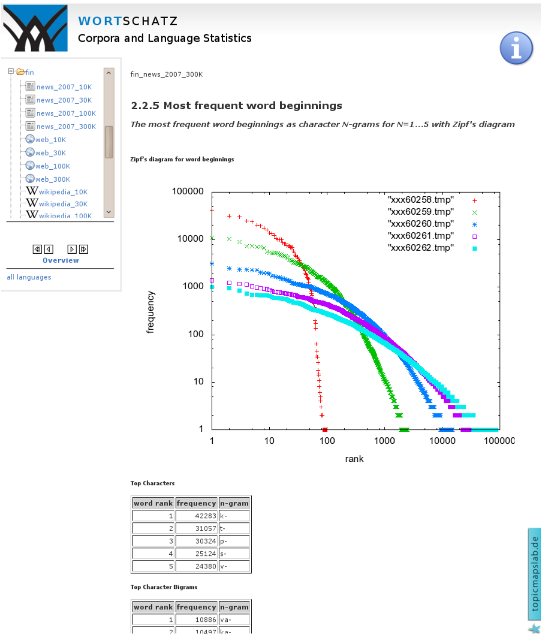
These sentences that were segmented by the sentence boundary detection and accepted by the following quality assurance procedures include “*Територією області течуть річки Ілі з притоками Чарин, Чілія, Текес, Курти; Каратал із притокою Коксу; Аксу; Лепси; Аягуз; Тентек; Кеген.*” (Ukrainian), “*Machiaj: Divizia Make-up DUMAREX Parteneri media: EVENIMENTUL ZILEI; ZIUA; JURNALUL NATIONAL; CAPITALA; COTIDIANUL; METROBUS; AZI; BURDA ROMANIA; ANTENA 1 - Doina Levintza, "'Neata"; PRIMA TV - "Clubul de Duminica", "Stil".*” (Romanian) or “*Siippainen kirjoitti lehtijuttujaan eri nimimerkeillä kuten Iloinen, Petteri; Kaaleppi; Karho, Otto; Kimpinen; Kimpinen, Kalle; Mäikiä, Urmas; O. S.; O. S-nen; Robin Hood; Saarto, Olavi; Svejk; Uolevi.*” (Finnish).

This information provides a fast feedback and leads to more accurate data resources in the future. Statistical values that may indicate problems with input selection, inaccurate preprocessing tools or other issues are widely spread, ranging from character analysis to show character set problems to automated rating of the corpora sources based on their homogeneity of various statistical values. This is still to be evaluated.

6. Presentation of the Results

Central goal for the presentation of the created result pages was a web portal that should allow both researchers in the fields of natural language processing and linguistics an easy access and overview of existing corpora and a starting point for evaluating linguistic phenomena in the field of corpus, genre and language comparison.

Each question, answered for a certain corpus, produces an HTML page containing the results. As described above, these result pages consist of a plot or of a (set of) table(s), or both. For comparisons, all corpora are assigned to three different categorization dimensions: language, text genre and corpus size. The Corpora and Language Statistics Website presented at *www.cls.informatik.uni-leipzig.de* supports this complex navigation. To achieve an easy access, despite the thousands of pages strongly related to each other, the ISO standard Topic Maps was used as underlying technology. Based on JRuby Topic Maps (cf. Bleier et al., 2009) and tinyTIM, all existing resources were merged while allowing extensions to new fields and dimensions in the future.



fin_news_2007_300K

2.2.5 Most frequent word beginnings

The most frequent word beginnings as character N-grams for N=1...5 with Zipf's diagram

Zipf's diagram for word beginnings

Top Characters

word rank	frequency	n-gram
1	42283	k-
2	31057	t-
3	30324	p-
4	25124	s-
5	24380	v-

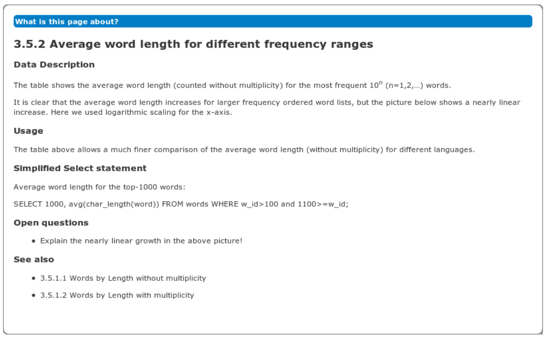
Top Character Bigrams

word rank	frequency	n-gram
1	10886	va-
2	10497	ka-

Figure 3: Sample HTML page

The user interface is designed to lower the entry barrier for the (possibly inexperienced) user: on the left side one can select between other languages, genres and corpus sizes. These links will show the corresponding page for the same question, but another corpus. The arrows allow linear scrolling through the different questions for one corpus.

An additional help screen gives detailed information about the data shown and the intentional background of the question. A (possibly slightly simplified) select-statement is provided. This can be used or modified for similar questions asked by the user. Some open problems and cross references complete this help screen.



What is this page about?

3.5.2 Average word length for different frequency ranges

Data Description

The table shows the average word length (counted without multiplicity) for the most frequent 10^n (n=1,2,...) words.

It is clear that the average word length increases for larger frequency ordered word lists, but the picture below shows a nearly linear increase. Here we used logarithmic scaling for the x-axis.

Usage

The table above allows a much finer comparison of the average word length (without multiplicity) for different languages.

Simplified Select statement

Average word length for the top-1000 words:

```
SELECT 1000, avg(char_length(word)) FROM words WHERE w_id>100 and 1100>=w_id;
```

Open questions

- Explain the nearly linear growth in the above picture!

See also

- 3.5.1.1 Words by Length without multiplicity
- 3.5.1.2 Words by Length with multiplicity

Figure 4: Sample help screen

7. Experimental Setup

7.1 Configuration of a Single Analysis

To allow contributions by different kind of persons including undergraduate students of different disciplines the underlying XML schema was designed in an uncomplex way that is nonetheless powerful through its universality.

There already exists a huge amount of scripts in different analysis domains. Therefore the standard procedure to extend the stock of evaluations is the modification of a template or an already used script and the adaption to the new problem.

Every task (as requests to the database management system, further processing like linking of temporary results or defining the specific visual output) is a single working step. As most new scripts try to examine an already considered field in more detail most parts of an existing script are still valid and can be adopted (especially simple post processing or output definitions). Therefore the effort of further extension is quite low. As a consequence whole ranges of new scripts could be generated by very simple replacements in already used SQL statements and explanatory text strings.

Listing 1 shows an excerpt of a simple evaluation script with all changes highlighted that are necessary to adapt the script to a new character.

```
<title>Distribution of Letter F</title>
<description>Number of sentences containing a fixed
          number of occurrences of this
          character</description>

<step descriptor="0">
          <sql-step>
                    <statement>select
          round(char_length(sentence)-
          char_length(replace(lower(sentence),"f","")))
          as freq, count(*), sentence from
          BASEDB.sentences group by freq order by
          freq</statement>
           </sql-step>
 </step>
```

Listing 1: Excerpt of an evaluation script

7.2 Reproducibility of the Results

To compare results of the CLS Website with similar results on other corpora it is essential to have free access to the corpora used here. Moreover, it must be transparent how the measurement was performed. The first condition is fulfilled by the availability of the Leipzig corpora collection, the second by the detailed description given in the help screens.

8. Further Work

At present not all existing corpora are already evaluated, many are still to be processed. To enhance usability and to achieve an easier access to the evaluation data it is intended to offer more interactive ways in the future. These will allow the user to compare values across self-chosen corpora and to inspect the data in more detail. Another aim is the adoption of the created tools and structures to other domains. As an example in eAQUA (cf. Heyer &. Schubert, 2008), a co-operational project of researchers of Computer Science and Ancient Science, a similar approach is used to give both sides a fast comparison of existing data resources and helps finding problems in the complex (pre-)processing of ancient texts.

9. References

Bleier, A.; Bock, B., Schulze, U., Maicher, L. (2009): *JRuby Topic Maps*. In *Proceedings of the Fifth International Conference on Topic Maps Research and Applications* (TMRA 2009). Leipzig, Germany.

Dickinson, M. and Meurers, D. (2005): *Detecting Annotation Errors in Spoken Language Corpora.* In: *Proceedings of the Special Session on Treebanks for Spoken Language and Discourse at the 15th Nordic Conference of Computational Linguistic* (NODALIDA-05), Joensuu, Finland.

Dunning, T. (1993): *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, Volume 19, number 1.

Eskin, E. (2000): *Automatic Corpus Correction with Anomaly Detection*. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-00). Seattle, Washington, USA.

Heyer, G. and Schubert, C. (2008): *eAQUA - Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft*. Word Wide Web electronic publication, http://www.eaqua.net.

Quasthoff, U.; Richter, M.; Biemann, C. (2006): *Corpus Portal for Search in Monolingual Corpora.* In: *Proceedings of the fifth international conference on Language Resources and Evaluation*, LREC 2006, Genoa, Italy.

Quasthoff, U. and Eckart, T. (2009): *Corpus build process of the project 'Deutscher Wortschatz'.* Workshop “Linguistic Processing Pipelines“, GSCL Conference 2009, Potsdam, Germany.

Zipf, G.K. (1949): *Human behavior and the principle of least effort : an introduction to human ecology*. Hafner reprint, New York, 1972, 1st ed. (Addison-Wesley, Cambridge, MA, 1949).