The Influence of Corpus Quality on Statistical Measurements on Language Resources

Thomas Eckart, Uwe Quasthoff, Dirk Goldhahn

Natural Language Processing Group, University of Leipzig, Germany
Johannisgasse 26, 04103 Leipzig
{teckart, dgoldhahn, quasthoff}@informatik.uni-leipzig.de

Abstract

The quality of statistical measurements on corpora is strongly related to a strict definition of the measuring process and to corpus quality. In the case of multiple result inspections, an exact measurement of previously specified parameters ensures compatibility of the different measurements performed by different researchers on possibly different objects. Hence, the comparison of different values requires an exact description of the measuring process. To illustrate this correlation the influence of different definitions for the concepts *word* and *sentence* is shown for several properties of large text corpora. It is also shown that corpus pre-processing strongly influences corpus size and quality as well. As an example near duplicate sentences are identified as source of many statistical irregularities. The problem of strongly varying results especially holds for Web corpora with a large set of pre-processing steps. Here, a well-defined and language independent pre-processing is indispensable for language comparison based on measured values. Conversely, irregularities found in such measurements are often a result of poor pre-processing and therefore such measurements can help to improve corpus quality.

Keywords: Corpus quality, Standardization, Statistical Evaluation

1. Introduction

In general, the quality of a result can be described either by a specification of the process creating this result or by inspection of the result. In the case of multiple result inspections, an exact measurement of previously specified parameters ensures compatibility of the different measurements performed by different researchers on possibly different objects. Hence, the comparison of different measured values requires an exact description of the measuring process.
Today, even simple measurements like size of a corpus measured by the number of tokens or the average sentence length measured in words are not comparable. In this paper, the effects of different specifications of the measured objects are shown for different corpora and languages. Different definitions for the measured objects are discussed.
All specifications are language independent. Hence, the values measured for different languages can be used for language comparison.

2. Different Quantities to be measured

In this paper, the following quantities describing a corpus will be considered in more detail:

- Number of tokens
- Type-Token-Ratio
- Average word length
- Average sentence length (measured in number of words)

There is always the question whether the values measured for a corpus can be generalized as being valid for the corresponding language. The values measured may depend on the text genre or the corpus size.

3. The Influence of Pre-processing

Especially for Web corpora, an extensive pre-processing is used in the corpus building process. Possible measurements are not performed directly on the HTML files, but on the text extracted from these pages. This pre-processing will influence the values to be measured later. Assume the task is to build a Web corpus in a given language. The following steps are usually performed:

1. Web Crawling: Collect HTML-Pages containing text assumed to be in the specific language. Usually only a random sample of all such texts is collected. This randomization does not affect the measurements very much. Usually the collection contains texts in other languages which will be removed in Step 3.
2. HTML-Stripping: Remove all HTML-code and additional markup, leaving plain text, but also boilerplates (Baroni et al., 2008).
3. Text cleaning: Some of the following steps are optional and are not considered relevant by some corpus builders (Quasthoff and Eckart, 2009):
   (a) Boilerplate removal
   (b) Removal of foreign language parts (whole texts or sentences)
   (c) (Optional) Removal of parts which are not well-formed sentences, using pattern matching methods
   (d) (Optional) Removal of duplicate sentences
   (e) (Optional) Removal of near duplicate sentences (see below).

Unfortunately, all these steps influence the measured values in the forthcoming measurement process.

4. The Influence of Standardization: Defining Sentences and Words

As a foundation for counting objects one needs a definition of the objects to be counted. The situation is difficult because there is no generally accepted linguistic definition for the concepts sentence or word. For calculating the number of sentences, the definition of the concept sentence is implicitly implemented in any sentence segmentation tool. For different good sentence segmentation tools, the number of segments a text is split into should not differ very much and can be taken as starting point for the number of sentences. However, the optional steps 3c and 3d above can reduce this number substantially. Experiences have shown that in extensive Web crawling, a reduction by 50% is realistic and can even be higher for specific genres. Especially when relying on newspaper sources with large textual overlaps (different versions of an article over time, news agency articles used by several newspapers etc.) reduction of up to 65% of the original set of sentences was reached.
To illustrate this fact table 1 shows the number of sentences after different stages of pre-processing. The second column shows the number of all identified sentences based on the input material, column 3 the number of sentences after some pattern-based cleaning procedures to eliminate obvious non-sentences and column 4 the number of sentences after additional removal of all duplicates.

Corpus	Number of all identified sentences	After pattern-based cleaning procedures	After duplicate removal
Korean	20,574,734 (100%)	15,556,544 (75.6%)	4,163,599 (20.2%)
Lombard	151,897 (100%)	87,843 (57.8%)	33,881 (22.3%)
Persian	2,222,623 (100%)	1,143,295 (51.4%)	670,599 (30.2%)

Table 1: Number of sentences after different pre-processing steps for exemplary corpora of the Leipzig Corpora Collection

The case of words is even more complicated (Fuhrhop, 2008) (ISO 24614-1, 2010). Corpus linguists may agree on a definition based on character strings surrounded by white spaces or punctuation marks. The number of such words may not differ very much for slightly different specifications. But the optional (though important) next step is to remove all non-words. This includes nonsense strings, but maybe also numbers, URLs, obvious typing errors etc. For automatic processing, there is usually a pattern-based description of non-words. Due to poor standardization efforts, these patterns differ very much. Hence, the number of words differs as well.
An analysis based on three corpora in different languages shows this correlation of strictness of pre-processing and number of gained word types. Table 2 shows the number of identified words for three different word definitions and therefore three different cleaning procedures. Column 2 states the number of all types identified by a standard tokenization procedure, for column 3 all types were removed that contained characters not in a specified set of letters, numbers and some special characters. The last cleaning procedures (column 4) additionally remove words containing numbers and other ill-formed terms based on set of patterns. All percentages are calculated regarding the number of all primarily identified types (column 2).
Apparently these three different word definitions can cause a loss of more than half of all identified types and lead therefore to completely different results in comparative studies. The stronger elimination of (mostly infrequent) types for the English corpus is due to the extremely heterogeneous input material.

Corpus	Number of all types	Number of types with only valid characters	Additional cleaning procedures
English	53,326,503 (100%)	29,546,612 (55.4%)	23,565,145 (44.2%)
French	6,947,779 (100%)	5,498,258 (79.1%)	4,900,502 (70.5%)
Icelandic	6,706,387 (100%)	5,941,986 (88.6%)	5,337,584 (79.6%)

Table 2: Number of types for three different word definitions and different corpora

5. The Effect on Measured Values

If the non-words identified in the above section are removed, the number of words decreases, too. This has effects on several quantities related to this number and will be shown for three different corpora:

- an English corpus consisting of newspaper and Wikipedia articles and Web pages with around 815 million sentences,
- a French corpus consisting of newspaper and Wikipedia articles and Web pages with around 75 million sentences,
- an Icelandic corpus consisting of Web pages and Wikipedia articles with around 38 million sentences.

5.1. Number of Tokens

As a consequence of the reduction of types the number of tokens of a corpus decreases too, in some cases dramatically. Table 2 shows these developments with a loss of tokens up to 5 percent.

5.2. Type-Token-Ratio

A popular measure for the analysis of text or corpora is the type-token-ratio. For the following values the simple definition $TTR = |Number\ of\ types|/|Number\ of\ tokens|$ was used. Table 4 shows the development of the type-token-ratio for the three used word definitions. The cleaning procedures remove mostly infrequent words. Hence, the TTR decreases.

Corpus	Number of all tokens	After cleaning procedure 1	After cleaning procedure 2
English	14,529 (100%)	14,403 (99.1%)	14,138 (97.3%)
French	1,449 (100%)	1,440 (99.4%)	1,377 (95.0%)
Icelandic	564.2 (100%)	558.8 (99.0%)	537.7 (95.3%)

Table 3: Number of tokens (in million) for three different word definitions and different corpora

Corpus	TTR without cleaning	After cleaning procedure 1	After cleaning procedure 2
English	0.0037	0.0021	0.0017
French	0.0048	0.0038	0.0036
Icelandic	0.0119	0.0106	0.0099

Table 4: Type-Token-Ratio for word lists after different cleaning procedures

5.3. Average Word Length

Table 5 shows the development of the average word length. Apparently, the restriction on a set of valid characters (column 3) leads in some cases to a reduction of the average word length. The reason for this reduction is mainly due to the elimination of HTML/JavaScript-Markup (that remained after the HTML extraction), URLs and email addresses and due to the elimination of enumerations like "Europe/Moyen-Orient/Afrique" or "electrical/mechanical". As the second cleaning procedure primarily deals with the elimination of (often short) numbers, the average word length increases in column 4.

Corpus	Average word length for all types	After cleaning procedure 1	After cleaning procedure 2
English	12.08	10.13	10.53
French	9.64	9.30	9.50
Icelandic	11.49	11.51	11.75

Table 5: Average word length for word lists for different cleaning procedures

5.4. Average Sentence Length

Table 6 shows the development of the average sentence length measured in number of words. As the number of sentences is unchanged, the average length decreases with a decreasing number of tokens.

6. Evaluation of the Results: Distribution of Measured Values as Corpus Quality

Measured values are the basis for a variety of different analysis. In many cases, the distribution of the measured values is expected to follow some well-known probability distribution. On the contrary, if the measured distribution does not follow the expectations, this might be an indication of poor pre-processing, especially in steps 3c) and 3d), or of overall poor corpus quality (Eckart et al., 2012).
As an example, figures 1 and 2 show the sentence length distributions for two corpora (measured in characters).

Corpus	Average word length for all types	After cleaning procedure 1	After cleaning procedure 2
English	17.83	17.67	17.35
French	19.37	19.25	18.41
Icelandic	17.21	17.04	16.40

Table 6: Average sentence length for different corpora after different cleaning procedures

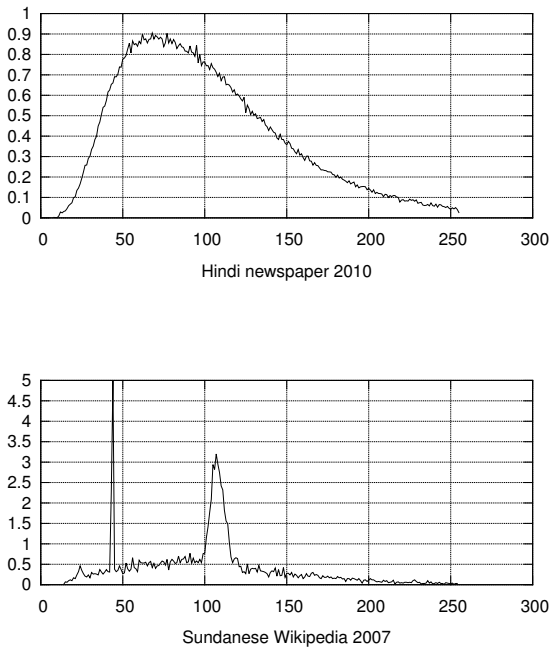

Figure 1: Sentence length distribution for two corpora (percentage for number of characters)

Apparently the latter (based on the Sundanese Wikipedia) shows a deformation compared to the typical distribution (here: a Hindi newspaper corpus) which can be found in corpora for most languages. In this case two peaks stand out: the first peak is due to a large set of nearly identical sentences that were not removed by duplicate detection:

- Taun ka-1118 Maséhi dina Kaléndar Grégorian.
- Taun ka-1119 Maséhi dina Kaléndar Grégorian.
- Taun ka-1120 Maséhi dina Kaléndar Grégorian.
- (English translation: This article is about the year X of the Gregorian calendar.)

The second peak is due to the following kind of sentences:

- Ancol nyaéta salasahiji désa di kacamatan Cinéam, Kabupatén Tasikmalaya, Propinsi Jawa Barat, Indonésia.
- Babakan nyaéta salasahiji désa di kacamatan Wanayasa, Kabupatén Purwakarta, Propinsi Jawa Barat, Indonésia.
- Bakung Lor nyaéta salasahiji désa di kacamatan Klangenan, Kabupatén Cirebon, Propinsi Jawa Barat, Indonésia.
- (English translation: X is a village in district Y, regency Z of the West Java Province, Indonesia.)

This second kind of near-duplicates shows much more variability and therefore, is harder to detect.
If one has to create a corpus for reliable statistical measurements, any irregularity in a distribution might be cured using a better pre-processing. As seen in the examples above, near duplicate sentences are responsible for several irregularities. Hence, near duplicate detection is crucial for reliable statistics. Of course, the term near duplicate sentences has to be defined in an exact way to be agreed upon. The following remarks can give a starting point.

- If two sentences are near duplicate, they share most of their words. The exact threshold has to be specified.
- If two sentences are near duplicate, they are of similar length.
- In many (but not all) cases they have the same beginning and/or end.

It is of interest to distinguish between small sets of near duplicate sentences and larger clusters. They differ by origin, need different detection methods and have different influence on the measured values. For small sets of near duplicate sentences we find:

- Pairs (or other small sets) of near duplicate sentences often differ by punctuation and use different types of quotation marks.
- Simple pattern-based rules help to reduce the number of such near duplicate sentences.
- The influence of small sets of near duplicate sentences on the measured values is usually small.

In contrast, for larger sets of pairwise near duplicate sentences:

- Larger sets of pairwise near duplicate sentences often differ in a single number (example: daily prices) or one (or a few) words as the examples above. They can be considered as variables and usually produce many near duplicate sentences.
- Bigger sets of near duplicate sentences are easier to detect because they can be found using clustering algorithms.
- Larger sets of pairwise near duplicate sentences are of greater importance, because they bias the statistical results. For example, larger sets of pairwise near duplicate sentences cause unexpected significance for word co-occurrences.

7. Conclusion

The value of statistical measurements strongly depends on their reproducibility and comparability. Even small changes in used definitions or working steps can lead to uncomparable and unappraisable results. This especially holds for Web corpora with a large set of pre-processing steps. Here, a well-defined and language independent pre-processing is indispensable for language comparison based on measured values. Conversely, irregularities found in such measurements are often a result of poor pre-processing and therefore such measurements can help to improve corpus quality.

8. References

M. Baroni, F. Chantree, A. Kilgarriff, and S. Sharoff. 2008. Cleaneval: a competition for cleaning web pages. In *Proceedings of the Sixth Language Resources and Evaluation Conference*, Marrakech.

T. Eckart, U. Quasthoff, and D. Goldhahn. 2012. Language statistics-based quality assurance for large corpora. In *Proceedings of Asia Pacific Corpus Linguistics Conference*, Auckland, New Zealand.

N. Fuhrhop. 2008. *Das graphematische Wort (im Deutschen): Eine erste Annaeherung*. Walter de Gruyter, Zeitschrift fuer Sprachwissenschaft, Vol. 27 edition.

ISO 24614-1. 2010. Language resource management – Word segmentation of written texts – Part 1: Basic concepts and general principles. ISO, Geneva, Switzerland.

U. Quasthoff and T. Eckart. 2009. Corpus Building Process of the Project 'Deutscher Wortschatz'. In *Linguistic Processing Pipelines Workshop at GSCL*, Potsdam, Germany.