How large is the BNC? A proposal for standardised tokenization and word counting

Vaclav Brezina and Matt Timperley (Lancaster University, UK)

Introduction

One of the core principles of the scientific method is replicability of results. Observations and experiments need to be exactly repeatable with different datasets in order to establish the stability of the findings (e.g. Asendorpf, 2013). Corpus linguistics aspires to being a scientific approach to language analysis and thus needs to satisfy the replicability requirement. This primarily involves standardisation of the procedures and instruments used in the field (McEnery & Hardy 2011).

Replicability in corpus linguistics is a fundamental issue which, however, has received only limited attention. For example, the differences in language use between multiple corpora sampling the same type of language have been explored in Gablasova et al. (2017) showing a large amount of variation between these corpora. We aim to take this research one step further to investigate the effect of different instruments with the same dataset. As an example dataset, we used the British National Corpus (BNC), which is a widely used dataset in corpus linguistic research, and six corpus tools commonly used for the analysis of the BNC.

In this experiment, we observe both i) variation in the overall token counts given by different tokenization procedures implemented by the tools as well as ii) variation in the frequency counts of individual linguistic variables, which is also connected to different ways of identifying tokens in the corpus. The main purpose of the experiment is to empirically analyse the amount of variation observed and assess its impact on the replicability of results in corpus linguistic studies. It is important to realise that the variation that we observe in this study is purely methodological and results from employing different tools. This type of variation is thus highly problematic because it shows that results produced by different tools are not comparable. The results therefore cannot be used easily in e.g. meta-studies, which bring together the knowledge in the field.

As mentioned earlier, the source of the variation in our experiment is the tokenization procedure and the lack of standardisation in this area. Tokenization is the process of splitting text into atomic parts. In segmented languages, like English, this is often considered an easy task because it can rely on graphical clues in the texts (spaces, punctuation etc.). This gives rise to a so called 'graphic word' as a unit (atomic part). Kučera and Francis (1967: 3) define a graphic word as "a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes but no other punctuation marks". The graphic word, however, is not the only definition employed in tokenizers for corpus tools; additional tokenization criteria are applied to deal with particular languages and ambiguous cases. These are, however, rarely openly stated.

The difficulties with tokenization generally arise in the following cases:

1. **Non-segmented languages.** Languages such as Chinese do not have delimiters to denote separation between words. This needs to be added using a language-specific process called segmentation.
2. **Punctuation.** Different dashes/hyphens can indicate that two strings of characters should be considered one, or, in other cases, two units, e.g. a forget-me-nots vs. a word-a string of characters-which… An apostrophe can indicate the end of an open quote, possessive case (e.g. Peter's) or a contraction (C'mas).
3. **Clitics.** Different decisions can be made about clitics and their status. Clitics such as 'll can be treated as atomic units or can be considered a part of larger atoms like we'll.
4. **Abbreviations.** Full stops may not occur with graphic words but might be considered an atomic unit when part of an abbreviation e.g. in this example.
5. **Multi-word expressions.** This overlaps with clitics. For some purposes, idioms or other multi-word expressions may be considered atomic (Webster and Kit, 1992).
6. **Data noise.** If the corpus texts were obtained via e.g. Optical Character Recognition (OCR) then characters may be mistaken for others. This can lead to situations that violate a static definition of an atomic unit.

The difficulties of tokenization are well presented in Yamashita and Matsumoto (2000), who propose a method of treating segmented and non-segmented languages with a single tokenizer. However, currently, there is no single tokenizer, or set of identical principles, used for all corpus tools. This means that when counting tokens there will be discrepancies between tools. The ambiguities might skew the counts in only a small proportion of cases, but this can be inflated by corpus size. We continue by investigating the impact of decisions taken at the tokenization stage on the analysis of the BNC. 3

Method

The BNC XML version, one identical dataset, was used in different corpus tools. The impact of different tokenization procedures on the searches was explored. Six different software tools were used**:** CQPWeb, BNCweb, BNC-BYU, Sketch Engine, Xaira and #LancBox.

Results

The results show that the overall token counts for the same dataset, the BNC, vary from 96,263,399 in the BYU interface to 112,289,776 in the Sketch Engine. The main source of the variation is the decision to include or exclude punctuation as token counts.

	CPQWeb	BNCweb	BYU	SkE	Xaira	LancBox
Tokens used for normalisation	112, 102, 325	98, 313, 429	96, 263, 399	112, 289, 776/ 112, 181, 015 (CLAWS)	112, 532, 992	96, 960, 485
Words (if different				96, 133, 793/ 96, 052, 598		

counts provided)						

Table 1. Size of the BNC in different corpus tools

Tokenization does not only affect the total word counts in the corpus and normalization of the data, it also plays a role in searches for individual linguistic variables as is apparent from Table 2 below.

Tools / Ling. variable	CPQWeb	BNCweb	BYU	SkE	Xaira	LancBox
the	6,041,234	6,041,234	5,971,799	6,054,939	6,055,159	6,054,559
the per 1M	53,890.35	61,448.72	62,036.03	53,922.40	53,807.86	62,443.57
new	124,022	124,022	121,881	124,399	124,308	124,235
new per 1M	1,106.33	1,261.5	1,266.12	1,107.80	1,104.64	1,281.30
research	26,682	26,682	26,566	26,702	26,793	26,692
research per 1M	238.01	271.4	275.97	237.80	238.09	275.29
mauve	214	214	213	208	214	206
mauve per 1M	1.91	2.18	2.21	1.85	1.90	2.12

Table 2. Frequencies of linguistic variables in the BNC according to different tools

Focusing on the raw frequencies (first entry row for each linguistic variable), we can see that in the case of the, the most frequent lexical item in English, the frequency counts range from 5,971,799 in the BYU platform to 6,055,159 in Xaira; this is a 1.4% difference. In the case of new and research, two mid-frequency items, the differences between the smallest and the largest frequency count are 2.07% and 0.85% respectively. Finally, the low-frequency item mauve shows a difference of 3.9%. The relative (normalised) frequencies (counts per 1M tokens), which combine the effect of total token counts for the BNC in the individual tools and the frequency counts for the individual variables, indicate even larger discrepancies among the individual tools.

Conclusion: Proposal for more rigorous analyses

Having shown the implications of different tokenization principles implemented in different widely-used corpus tools and having demonstrated the effect of these on the results of quantitative analyses in corpus linguistics, we propose two innovations: i) Tokenization Parameters Notation (TPN) and ii) LOB as a unit of corpus size measurement. First, we introduce the Tokenization Parameters Notation (TPN), which uniquely describes the principles of tokenization implemented in a corpus tool. Second, we introduce LOB, a new unit of corpus measurement and normalization, which is independent of the individual corpus tokenization. LOB is defined as the size of the Lancaster-Oslo/Bergen corpus (LOB) given the particular tokenization principles. It is approximately one million tokens but can vary across different corpus tools. The advantage of using LOB-based normalization, as opposed

to normalization to a particular fixed basis, e.g. per million, is that it allows larger comparability of results based on different tools. 5

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K. & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2), 108-119.

Kučera, H. & Francis, W.N. (1967). *Computational Analysis of Present-Day American English*. Brown University Press.

Gablasova, D., Brezina, V. & McEnery, T. (2017, forthcoming). *Exploring learner language through corpora: comparing and interpreting corpus frequency information. Language Learning*.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on Computational linguistics-Volume 4* (pp. 1106-1110). Association for Computational Linguistics.

Yamashita, T., & Matsumoto, Y. (2000). Language independent morphological analysis. In *Proceedings of the sixth conference on Applied natural language processing* (pp. 232-238). Association for Computational Linguistics.