

The TenTen Corpus Family

Miloš Jakubíček^{▲▼}, Adam Kilgarriff[▲],
Vojtěch Kovář^{▲▼}, Pavel Rychlý^{▲▼},
Vít Suchomel^{▲▼}

▲ Lexical Computing Ltd., United
Kingdom

▼ Masaryk University, Czech Republic

<name>.<surname>@sketchengine.co.uk

Introduction

Everyone working on general language would like their corpus to be bigger, wider-coverage, cleaner, duplicate-free, and with richer metadata. In this paper we describe our programme to build ever better corpora along these lines for all of the world's major languages (plus some others).

Baroni and Kilgarriff (2006), Sharoff (2006), Baroni et al (2009), and Kilgarriff et al (2010) present the case for web corpora and programmes in which a number of them have been developed. TenTens are a development from them.

Names

Two of the programmes above used the WaC suffix for corpus-naming. To forestall confusion with a name like FrWaC being ambiguous between two different corpora (though both French and web-crawled) a new name was needed. The new batch of corpora are in the order 10^{10} (10 billion) words, so this is the TenTen family.¹ The corpus name is then formed by prefixing with the two-letter ISO-639-1 code for the language, and, optionally, suffixing with two-digits for the year of collection, to give e.g. enTenTen12 for English collected in 2012, zhTenTen for Chinese.

¹ We continue to use the WaC suffix in the 'Corpus Factory' programme, which uses slightly different methods (see Kilgarriff et al. 2010), mainly for languages with fewer speakers and less of a web presence.

Major world languages

We treat the following as major world languages (based on number of speakers and sizes of associated economies): Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish. We have created, and will maintain and develop, TenTen corpora for each of these eleven languages. We have also developed them for several other languages we have particular interests in, currently Czech, Hungarian, Polish and Slovak.

All these corpora are available within the Sketch Engine (Kilgarriff et al 2004).²

Spiderling, jusText, Onion

The processing chain for creating the corpus is:

- Crawl the web with spiderling³ (Pomikalek and Suchomel 2012), a crawler designed specifically for preparing linguistic corpora
- Remove non-textual material and boilerplate with jusText (Pomikalek 2011). JusText uses the working definition that we want only 'text in sentences' (and not, e.g. headers and footers). The algorithm is linguistically informed, rejecting material that does not have a high proportion of tokens that are the grammar words of the language, so, in the course of data-cleaning, most material which is not in the desired language is removed.
- De-duplicate with onion (Pomikalek 2011). We de-duplicate at the paragraph level, as, for many linguistic purposes, a sentence is too small a unit, but a whole web page (which may contain large chunks of quoted material) is too large.

These tools are designed for speed and we use them installed in a cluster of servers. For a language where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day. The 12-billion-word enTenTen12 was collected, in 2012, in twelve days.

² <http://www.sketchengine.co.uk>

³ <http://nlp.fi.muni.cz/trac/spiderling>

Then, we want to tokenize the corpus into words, lemmatise, and part-of-speech tag. For these processes we examine the available tools for the language and apply the best we can find (after considering, firstly, accuracy, but also speed, quality of engineering, and licence terms). We have made extensive use of TreeTagger and FreeLing for European languages; Stanford tools for Chinese, meCab (with UniDic lexicon) for Japanese, Han Nanum for Korean, and MADA (in collaboration with Columbia University) for Arabic.

Static corpora and monitor corpora

A static corpus is a fixed dataset. A monitor corpus moves on, adding more material over time, so it can monitor change in the language (Clear 1986). The advantage of the static corpus is that it is a fixed point that can be referred to in years to come and always means the same thing. The advantage of the monitor corpus is that it stays up to date.

We do not see these two goals as conflicting. Our plan is to re-crawl each language every year or two, and then, after filtering out any paragraphs in the new material that we already had in the old, adding the new to the old, with metadata that allows us to search in, and gather statistics over, ‘only the new’ or ‘only the old’. This also allows us to contrast the new with the old, using Sketch Engine functions such as keywords and sketch-diffs.

Virtual corpora

A corpus is a collection of texts. If you add one collection to another, you get a bigger collection. $1+1=1$. There are often benefits to treating two corpora of the same language as two parts of a larger whole. We have recently developed technology that implements the intuition, allowing two or more existing corpora, indexed in the Sketch Engine, to be seen as a single corpus from the user’s point of view.

Virtual corpora, or super-corpora, have several benefits. They make maintenance of these very large objects easier, as different component corpora can be stored and indexed separately. Also when we add new material, to a very large corpus, we will not need to re-index the whole. They encourage the super-corpus designer to be disciplined in their use of metadata fields, as queries will only make sense if

there is a unified system covering the metadata of all component corpora.

Fixed corpora: pros and cons

As already noted, many people would like their corpus to be fixed, so that queries and experiments run over it give exactly the same results now and in ten years time. Some argue that such replicability is central to the scientific integrity of the field.

This presents us with a substantial difficulty. We often find problems with our corpora, for example, sets of pages from a spam website. We would like to remove that spam, and the corpus will then be more useful for most users, but those who want replicable results object.

A similar issue arises with NLP tools. If there are better tools, or even just debugged or otherwise improved versions of those we are already using, should we upgrade? For most of our users, we would like to, but those who want replicability will object.

To some extent these problems can be solved by keeping numerous versions. But corpora are large, and management and maintenance is in any case a large task, and there are limits to our willingness to keep multiple versions.

As a policy, our priority is good, up-to-date data and mark-up, and we give higher priority to data quality than to 100% replicability. We think a metaphor from the natural sciences is more apt here than one from computer science. Where biologists replicate an experiment with a new sample of tissue, they do not expect 100% replicability. Replicability will be within margins according to the variability of the material under scrutiny.

Metadata

One of the limitations of web-crawled corpora is that they come with very little metadata.

Date of production is one problem: none of the dates on a web page reliably state when it was written – unless it is one of a few types of text such as newspaper, blog, or press release. We are supplementing general crawls (where we have the date of crawling, which is of some use, but little

else) with targeted crawls for these text types (see Minocha et al 2013).

Another concern is region. For Spanish, Portuguese and Arabic, we have metadata fields according to the top level domain of the website that the text came from. For English we have trained a classifier to distinguish British and American English, and applied it to all of enTenTen, so we have data-derived metadata.

We have also classified all documents in enTenTen for readability, based on Kilgarriff et al (2008) and plan to do the same for formality, using a method based on Heylighen and Dewaele (1999).

We are exploring domain corpora using both bottom-up methods and targeted crawling (Avinesh et al 2012) so in due course, large parts of the TenTen corpora will have a value for the 'domain' attribute.

Conclusion

We have presented a new family of corpora, the TenTens, of the order of 10 billion words. We have described how we are building them, what we have built so far, and how we shall continue maintaining them and keeping them up to date in the years ahead. While, as yet, they have very little metadata, we are working out how to gather and add metadata attribute by attribute. The corpora are all available for research at <http://www.sketchengine.co.uk>.

References

- Avinesh PVS, Diana McCarthy, Dominic Glennon and Jan Pomikálek (2012) Domain Specific Corpora from the Web Proc *EURALEX*. Oslo, Norway.
- M. Baroni and A. Kilgarriff. 2006. [Large linguistically-processed Web corpora for multiple](#)

- [languages](#). Conference Companion of EACL 2006.
- M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. [The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora](#). J. Language Resources and Evaluation 43 (3): 209-226.
- J. Clear. 1986. Trawling the language: Monitor corpora. Proceedings of Euralex.
- F. Heylighen and J-M Dewaele 1999. Formality of Language: definition, measurement and behavioural determinants. Internal Report, Free Univ Brussels.
- A. Kilgarriff, P. Rychly, P. Smrz, D. Tugwell. 2004. The Sketch Engine. Proc Euralex, Lorient, France.
- A. Kilgarriff. 2009. Simple Maths for Keywords. Proc Int Conf on Corpus Linguistics.
- A. Kilgarriff M Husak K McAdam M Rundell P. Rychly. 2008. GDEX: Automatically Finding Good Dictionary Examples. Proc. EURALEX, Barcelona.
- A. Kilgarriff, S. Reddy, J. Pomikalek, Avinesh PVS. 2010. A corpus factory for many languages. LREC, Malta.
- A. Minocha. S. Reddy and A. Kilgarriff 2013. Feed Corpus: an ever-growing up-to-date corpus. 8th Web-as-Corpus workshop, Lancaster, UK.
- J. Pomikalek 2011. Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis, Masaryk University, Brno, 2011.
- J. Pomikalek and V. Suchomel 2012. [Efficient Web Crawling for Large Text Corpora](#) Proc. 7th Web-as-Corpus workshop, Lyon, France.
- Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, (eds), WaCky! Working papers on the Web as Corpus. Gedit, Bologna.