Overview of Latin American and Iberian corpora in Sketch Engine

Michal Cukr[1], Miloš Jakubíček[1,3], Jan Kraus[1], František Kovařík[1,2], and Vít Suchomel[1,3]

[1] Lexical Computing, 602 00 Brno, Czech Republic
firstname.lastname@sketchengine.eu
[2] Faculty of Arts, Masaryk University, 602 00 Brno, Czech Republic
[3] Faculty of Informatics, Masaryk University, 620 00 Brno, Czech Republic

Abstract. This paper presents the biggest and most used corpora of Latin American and Iberian languages available in Sketch Engine. It provides insight into which corpora can be used for what purposes, which corpora are open to use and for which languages are the big corpora projects available at the time of writing.

Keywords: Catalan · Galician · Spanish · Portuguese · corpus · Sketch Engine · web corpora · parallel corpora · timestamped

Introduction

Sketch Engine is an online system used for text analysis and corpus management, largely used in but not restricted to the linguistic and lexicographic area.[7] It offers a vast set of tools, including Word Sketch – a simple-to-use analysis of typical collocations. Sketch Engine works with hundreds of large corpora in many languages, including Catalan, Galician, Portuguese, and Spanish. It lets users search and analyse corpora texts simply, and create corpora of their own (section 11)

This paper aims to present the most important corpora processed by Sketch Engine in said languages.

Corpus access

Some of the corpora presented in this paper are free to use without an account. A list is provided at `https://app.sketchengine.eu/#open`, from where these can be accessed.

Most public corpora in Sketch Engine require a registration before use. The first 30 days of a registered account are provided for free.

Sketch Engine enables the users to search corpora and perform language analysis. The access to source texts of corpora is subject to individual agreements.

1 The TenTen Corpus Family

The TenTen Family [6] is a group of some of the largest corpora today. The texts are gathered from the World Wide Web using advanced crawling techniques.[14] They aim for general language with representation among many text types. The most recent TenTen corpora are for Spanish esTenTen18 (17 billion words) [9], for Portuguese ptTenTen20 (12.6 billion words) [8], and for Catalan caTenTen14 (183 million words). All of these are free to use with at least a trial Sketch Engine account.

Sketch Engine allows users to work with subcorpora according to the national top-level domain. E.g. *.ar* is the Argentinian TLD. For esTenTen18, a Latin American subcorpus for all American Spanish countries and a subcorpus for each country are available. A Brazilian subcorpus is available for ptTenTen20.

The TenTen corpora, as well as Timestamped corpora (section 2), are useful for general language analysis, dictionary and language textbook making, designing brands and slogans for marketing, examples of typical use of a language, large language models etc.

Both the TenTen corpora and Timestamped web corpora are de-duplicated and cleaned from boilerplate (repetitive web page patterns, spam etc.) and texts in other languages. (More information about corpora cleaning is provided in [6].)

More information about esTenTen, ptTenTen and caTenTen can be found on the following websites:

- Spanish:
  `https://www.sketchengine.eu/estenten-spanish-corpus/`
- Portuguese:
  `https://www.sketchengine.eu/pttenten-portuguese-corpus/`
- Catalan:
  `https://www.sketchengine.eu/catenten-catalan-corpus/`

2 Trend Corpora

Timestamped web corpora (trend corpora) are made up of texts from newsfeeds [16] available in many languages (Catalan, Spanish, Portuguese) and are very large, often reaching billions of words.[4] Most texts come from news portals and lifestyle magazines. Every text provides information about the publishing date. Texts with the publication date until 2022 have been acquired by JSI. New texts are added weekly since 2023. Some of the Timestamped web corpora are free to use with at least a trial Sketch Engine account, other require a paid account.

Timestamped corpora are useful for diachronic analysis of new language trends (e.g. use of words/phrases in a timeline), hot topics etc.

3 Europarl Spoken Parallel and ParlaMint

The multilingual corpus Europarl consists of sentence-aligned corpora of 21 different languages of the EU, including Portuguese and Spanish.[10] The texts come from the official European Parliament proceedings. The Spanish part of the Europarl Spoken Parallel corpus is free to use without an account, others require at least a trial Sketch Engine account.

Similarly, ParlaMint parallel corpora consist of 17 multilingual comparable corpora of debates from European parliaments, including the Spanish parliament.[5]

4 OpenSubtitles 2018 Parallel

Another parallel corpus, OpenSubtitles was extracted from translated movie subtitles of `www.opensubtitles.com`. (Corpus: [15], extraction: [11].) It consists of 60 corpora in 58 languages, including Catalan, Galician, Spanish, European Portuguese, and Brazilian Portuguese. The use of all of these corpora requires a paid Sketch Engine account.

5 Gutenberg Corpora 2020

Project Gutenberg consists of 29 corpora made up of e-books from the Gutenberg database, with corpora of Catalan, Portuguese and Spanish.[2] The use of all of these corpora requires a paid Sketch Engine account.

6 CHILDES corpus

A comparable corpus of transcripts of child language. Catalan, Portuguese, and Spanish child language corpora are available. (More on the TalkBank web page: `childes.talkbank.org`.[1] Using the CHILDES corpus requires a paid account.

7 EUR-Lex

The EUR-Lex parallel corpus contains multilingual corpora from the EUR-Lex database.[3] The texts are public documents of the EU, including the European Union law and more. All the texts are translated into all 24 official languages of the EU. The corpus covers a large area of subjects and can be used as a general-purpose translation corpus.

8 DGT Translation Memory

DGT Translation Memory is a parallel corpus of aligned sentences in 24 languages, including Portuguese and Spanish. The texts come from the European Union's legislative documents (Acquis Communautaire) from a large translation memory DGT published by The European Commission.[13][12]

9 Brazilian Portuguese Corpus (Corpus Brasileiro)

Data for the Brazilian corpus have been gained from May 2008 to April 2010. It contains over a billion tokens, making it a large and useful resource for language analysis. It is rich in metadata, including detailed genre distribution (e.g. Journalism – Newspaper, Journalism – Magazine). The project has been funded by Fapesp (Sao Paulo State Research Foundation) and supported by CEPRIL (Center for Research and Information on Language), the Graduate Program in Applied Linguistics (LAEL) at Sao Paulo Catholic University (PUCSP), Brazil.

10 Drama Corpora (DraCor)

DraCor is a set of 21 corpora, containing drama texts in 14 languages. These are useful for research in digital humanities, literature studies, and linguistics. Available are the Spanish Calderón Drama Corpus with over 2 million words and the Spanish Drama Corpus with over 371 thousand words.

All drama corpora are free to use with at least a trial Sketch Engine account.

11 Users' corpora

The users can create and manage their own corpora directly in Sketch Engine. The users' corpora are processed and can be accessed with the same tools as the corpora above, including compilation, concordance search, Word Sketch, Thesaurus etc.

12 Lists of available corpora

More than 60 Catalan, Galician, Portuguese and Spanish corpora can be accessed in Sketch Engine. The following links lead to lists of corpora in said languages:

Spanish corpora:
`https://www.sketchengine.eu/corpora-and-languages/spanish-text-corpora/`
Portuguese corpora:
`https://www.sketchengine.eu/corpora-and-languages/portuguese-text-corpora/`
Catalan corpora:
`https://www.sketchengine.eu/corpora-and-languages/catalan-text-corpora/`
Galician corpora:
`https://www.sketchengine.eu/galician-text-corpora/`

Conclusion

We presented the corpora of Catalan, Galician, Portuguese and Spanish, currently available in the Sketch Engine tool. We included a brief description of the tool and each of the presented corpora – their purpose, basic properties, and their use.

References

1. CHILDES – child language corpus, https://www.sketchengine.eu/childes-corpora/, accessed on February 21, 2024
2. Gutenberg corpora 2020: 29 corpora of books, https://www.sketchengine.eu/gutenberg-corpora-2020/, accessed on February 21, 2024
3. Baisa, V., Michelfeit, J., Medveď, M., Jakubíček, M.: European Union language resources in Sketch Engine. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 2799–2803. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1445.pdf
4. Bušta, J., Jakubíček, M., Herman, O., Krek, S., Novak, B.: JSI Newsfeed Corpus. In: The 9th International Corpus Linguistics Conference. Corpus Linguistics 2017 Conference (07 2017), https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper382.pdf
5. Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L.D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Darģis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Bartolini, R., Cimino, A., Diwersy, S., Luxardo, G., Rayson, P.: Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1 (2021). https://doi.org/10.1007/s10579-021-09574-0, slovenian language resource repository CLARIN.SI
6. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. 7th International Corpus Linguistics Conference CL 2013 pp. 125–127 (07 2013), https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf
7. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1**, 7–36 (2014). https://doi.org/10.1007/s40607-014-0009-9
8. Kilgarriff, A., Jakubíček, M., Pomikálek, J., Sardinha, T.B., Whitelock, P.: PtTenTen: A corpus for Portuguese lexicography. In: Tony Berber Sardinha, T.d.L.S.B.F. (ed.) Working with Portuguese Corpora, pp. 280–287. Bloomsbury Publishing, London, 1 edn. (2014)

9. Kilgarriff, A., Renau, I.: esTenTen, a Vast Web Corpus of Peninsular and American Spanish. Procedia – Social and Behavioral Sciences **95**, 12–19 (2013). https://doi.org/10.1016/j.sbspro.2013.10.617
10. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of Machine Translation Summit X: Papers. pp. 79–86. Phuket, Thailand (08 2005), https://aclanthology.org/2005.mtsummit-papers.11.pdf
11. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 923–929. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1147.pdf
12. Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., Gilbro, S.: An overview of the European Union's highly multilingual parallel corpora. Language Resources and Evaluation **48**, 679–707 (12 2014). https://doi.org/10.1007/s10579-014-9277-0
13. Steinberger, R., Eisele, A., Klocek, S., Pilos, S., Schlüter, P.: DGT-TM: A freely available Translation Memory in 22 languages. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 454–459. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf
14. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Proceedings of the 7th Web as Corpus Workshop (WAC7). pp. 39–43 (04 2012), https://www.sketchengine.eu/wp-content/uploads/Efficient_Web_2012.pdf
15. Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12) (2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf
16. Trampus, M., Novak, B.: Internals of an Aggregated Web News Feed. In: Proceedings of 15th Multiconference on Information Society 2012 (IS-2012) (2012), https://aile3.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf