CIEP+

The Corpus of Indo-European Prose + more

Annemarie Verkerk & Luigi Talamo

September 2021

1 Introduction

In this document we describe CIEP+, the Corpus of Indo-European Prose Plus, which in fact samples non-Indo-European languages as well :). This corpus has a history that started in 2009, when Annemarie Verkerk started on her PhD at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. This was the start of a library collection of translations of *Alice's Adventures in Wonderland*, *Through the Looking-Glass and What Alice Found There* (Lewis Carroll), and *O Alquimista* (Paulo Coelho). A glossed corpus of extractions was compiled and this formed the main data set used by Verkerk (2014).

While the collection of translations never really stopped after that, the compilation of CIEP+ started for real when Luigi Talamo joined Saarland University / SFB 1102 at the end of 2019. The set of translations as well as the sample of languages was greatly expanded, driven by the sense of a need for such a resource in linguistic typology.

2 Parallel corpora

Parallel texts have been used as a resource for linguistic typology since the late nineties (Stolz and Gugeler 2000, see also Cysouw and Wälchli 2007; Wälchli and Cysouw 2012). They are described by Haspelmath (2018) as sources of token-based comparative concepts, units of cross-linguistic comparison rooted in translation contexts that are equivalent across originals and translations of the same work. There are many benefits of using parallel texts for typology (Wälchli 2007, p. 99), including 1) the possibility of including context or investigating contextual variation; 2) investigating language-internal variation; 3) given point 2, it becomes possible to obtain more sensitive typological measures. Parallel corpora also have disadvantages, mentioned prominently by Wälchli (2007, p. 132): "(a) written-language bias (Linell 1982), (b) bias toward planned (conscious) language use (including purism) (Miller & Weinert 1998), (c) bias toward religious and legalese registers, (d) narrative register bias, (e) bias toward large languages (in spread zones), (f) bias toward standardized (simplified?) language varieties, (g) bias toward

non-native use of languages, (h) bias toward translated language (rather than original language use)."

While we are aware of both benefits and disadvantages, with CIEP+ we try to fill a gap in the landscape of parallel corpora for typology. Famous usable parallel corpora in approximate order of appearance include:

- OPUS, the open parallel corpus by Jörg Tiedemann (Tiedemann 2012).[1] This includes EuroParl, a famous corpus containing the proceedings of the European Parliament in 21 European languages. Most of the many resources available on OPUS concern language pairs or relatively small samples. The exception is JW300, a parallel corpus containing articles from 'The Watchtower Announcing Jehovah's Kingdom' in over 300 languages (on average 100 thousand parallel sentences per language pair);

- InterCorp, a parallel corpus part of the Czech national Corpus, with coverage of 39 languages (Čermák 2012).[2] This is a compilation of different corpora (and different registers), many of which are not attested in all 39 languages. As is common, there is an emphasis on European languages;

- ParaSol, A Parallel Corpus of Slavic and other languages by Ruprecht von Waldenfels (Waldenfels 2011).[3] This is a parallel web-corpus of novels (like CIEP+), but with more literary works covered in a smaller language sample.

- Parallel Bible Corpus, a collection of Bible translations in over 1450 languages (Mayer and Cysouw 2014).[4] The Bible is the most widely translated text of humankind; and it is naturally parallel given retained verse numbers.

- The UDHR in Unicode project, has the purpose of showcasing Unicode in as many language of the world as possible, through translations of the Universal Declaration of Human Rights (UDHR).[5] Currently has translations in almost 500 languages.

- ParTy, the Parallel Corpus for Typology, contains automatically aligned subtitles of films in different languages from around the world Levshina 2016.[6] This is a unique corpus in that subtitles approximate spoken language far better than any of the resources listed above.

It is also relevant to mention several non-parallel original corpora, as some of these deal with spoken and/or conversational data that are very hard to access through parallel corpora, showing one of their weaknesses:

[1] `https://opus.nlpl.eu`
[2] `https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze9`
[3] `https://parasolcorpus.org`
[4] `https://github.com/cysouw/paralleltext/tree/master/bibles`
[5] `http://www.unicode.org/udhr/`
[6] `https://github.com/levshina/ParTy-1.0`

- Universal Dependencies treebanks (Marneffe et al. 2021);[7]
- DoReCo, Language Documentation Reference Corpora (first reference Paschen et al. 2020) contains over 50 languages, minimum corpus size is 10.000 tokens;[8]
- Multi-CAST, Multilingual Corpus of Annotated Spoken Texts[9] contains 15 languages, minimum corpus size is 1000 clauses (Haig and Schnell 2021);
- The DOBES Archive at The Language Archive, located at the MPI for Psycholinguistics;[10]
- The hzsk Collection, Hamburger Zentrum für Sprachkorpora;[11]
- Leipzig University's Corpora Collection.[12]

Several of these original corpora have are of spoken language use, most of them are characterized by a large disparity regarding the sampling of languages and by the materials that constitute each subcorpus.

3 CIEP+ books and languages

With CIEP+, we aim to fill a gap in the set of (parallel) corpora that are currently being used in typology. CIEP+ is a parallel corpus of originals and translations of 18 literay works (listed below) in 32 Indo-European and 10 non-Indo-European languages (see Table 1). We collect originals and translations of the following works:

- L. Carroll. *Alice's Adventures in Wonderland.* London: Macmillan, 1865;
- L. Carroll. *Through the Looking-Glass and What Alice Found There.* London: Macmillan, 1871;
- P. Coelho. *O Alquimista.* Rio de Janeiro: Rocco, 1989;
- P. Coelho. *O Zahir.* Rio de Janeiro: Rocco, 2005;
- U. Eco. *Il nome della rosa* Milano: Bompiani, 1980;
- A. Frank, O. Frank, M. Pressler. *Het Achterhuis: dagboekbrieven 12 juni 1942 - 1 augustus 1944.* Amsterdam: Uitgeverij Bert Bakker, 1947;
- N. Kazantzakis. *Víos kai Politeía tou Aléxē Zorbá.* Athens: Dēmētriou Dēmētrakou A.E., 1946;

[7] `universaldependencies.org`
[8] `http://doreco.info`
[9] `https://multicast.aspra.uni-bamberg.de`
[10] `https://archive.mpi.nl/tla/`
[11] `https://corpora.uni-hamburg.de/hzsk/en/repository-search`
[12] `https://corpora.uni-leipzig.de/en`

- G. García Márquez. *Cien Años de Soledad*. Buenos Aires: Editorial Sudamericana, 1967;
- G. Musso. *La jeune fille et la nuit*. Paris: Calmann-Lévy, 2018;
- J. K. Rowling. *Harry Potter and the Philosopher's Stone*. London: Bloomsbury Publishing Plc., 1997;
- J. K. Rowling. *Harry Potter and the Chamber of Screts*. London: Bloomsbury Publishing Plc., 1998;
- J. K. Rowling. *Harry Potter and the Prisoner of Azkaban*. London: Bloomsbury Publishing Plc., 1999;
- J. K. Rowling. *Harry Potter and the Goblet of Fire*. London: Bloomsbury Publishing Plc., 2000;
- J. K. Rowling. *Harry Potter and the Order of the Phoenix*. London: Bloomsbury Publishing Plc., 2003;
- J. K. Rowling. *Harry Potter and the Half-Blood Prince*. London: Bloomsbury Publishing Plc., 2005;
- J. K. Rowling. *Harry Potter and the Deathly Hallows*. London: Bloomsbury Publishing Plc., 2007;
- A. de Saint-Exupery. Le Petit Prince. Paris: Gallimard, 1943;
- P. Süskind. *Das Parfum: Die Geschichte eines Mörders*. Zürich: Diogenes, 1985.

Comparing CIEP+ with other resources that have been used in linguistic typology, it has some distinct benefits but also drawbacks. It is a parallel corpus, which implies that the content of the corpus are translational equivalents; i.e. the same content is contained and maintained across translations. This has a clear benefit for example over the Universal Dependencies treebanks (Marneffe et al. 2021), which are not parallel and often not even contain the same registers, which makes comparison more indirect. Second of all, we deal with the genre of fiction, which is relatively unexplored in corpus-based typology. Third, dealing with translations of highly popular novels ensures high data quality, as these were created and sold to make a profit. Fourth, the corpus contains both highly literary language in classics such as *Cien Años de Soledad* and *Das Parfum* as well as books with extensive dialog (*Harry Potter*, *Alice in Wonderland*) that allows both the investigation of language-internal variation and spoken-language-like qualities.

Of course, using CIEP+ for typological research also has drawbacks. The bulk of the corpus are translations, and might be effected by translationese, i.e. influences on the target text from the source text mediated by the translation process. The effects of using parallel corpora for typology are underexplored (but see Cappelle 2012). Another issue is the biased language sample; only relatively 'big' languages are included as only

these have translations of popular literary works. This means we will not be able to sample many (i.e. over a couple of hundred) languages, and CIEP+ has a genealogical and geographical bias on languages of Europe. We are exploring the options regarding alternative versions of CIEP, in which we only include (originals and) translations of the most widely translated books, *Le Petit Prince*, *Alice's Adventures in Wonderland*, and perhaps *Harry Potter and the Philosopher's Stone*.

Language	**IE?**	**(sub)family**	**books**	**w. order**	**SO ent.**	**status**	**UD parser**
Albanian	IE	Albanian	17	SVO	-	Spring 2022	less than 1K
Armenian	IE	Armenian	14	DC	-	Summer 2022	52K
Bulgarian	IE	Balto-Slavic	18	DC	0.36	almost finished	156K
Croatian-Serbian	IE	Balto-Slavic	18	SVO	0.47	completed	199K
Latvian	IE	Balto-Slavic	17	SVO	0.68	almost finished	252K
Lithuanian	IE	Balto-Slavic	18	SVO	0.94	almost finished	75K
Polish	IE	Balto-Slavic	18	SVO	0.70	completed	499K
Russian	IE	Balto-Slavic	18	SVO	0.58	completed	1,422K
Breton	IE	Celtic	4	SVO	-	almost finished	10K
Irish	IE	Celtic	5	VSO	0	almost finished	818K
Welsh	IE	Celtic	5	VSO	-	completed	36K
Danish	IE	Germanic	18	SVO	0.21	completed	100K
Dutch	IE	Germanic	18	mixed	0.31	completed	306K
English	IE	Germanic	18	SVO	0.07	completed	1,880K
German	IE	Germanic	18	mixed	0.23	completed	3,753K
Swedish	IE	Germanic	17	SVO	0.14	almost finished	206K
Modern Greek	IE	Hellenic	18	DC	0.36	completed	63K
Bengali	IE	Indo-Aryan	12?	SOV	-	Spring 2022	upcoming
Gujarati	IE	Indo-Aryan	5?	SOV	-	Spring 2022	no
Hindi	IE	Indo-Aryan	15?	DC	0.32	Spring 2022	375K
Marathi	IE	Indo-Aryan	11?	SOV	0	Spring 2022	3K
Nepali	IE	Indo-Aryan	6?	DC	-	?	upcoming
Punjabi	IE	Indo-Aryan	3?	SOV	-	Spring 2022	no
Sinhala	IE	Indo-Aryan	11?	SOV	-	?	upcoming
Urdu	IE	Indo-Aryan	8?	SOV	0.39	Spring 2022	138K
Persian	IE	Iranian	16?	SOV	0.22	?	654K
French	IE	Romance	18	SVO	0.03	completed	1,156K
Italian	IE	Romance	18	SVO	0.29	completed	818K
Latin	IE	Romance	6	mixed	0.58	almost finished	977K
Portuguese	IE	Romance	17	SVO	0.08	completed	571K
Romanian	IE	Romance	18	DC	0.18	almost finished	937K
Spanish	IE	Romance	18	SVO	0.19	completed	1,015K
Arabic	-	Afroasiatic	18	VSO	0.19	completed	1,042K
Basque	-	isolate	14	DC	0.75	Spring 2022	121K
Finnish	-	Uralic	17	DC	0.38	almost finished	397K
Georgian	-	Kartvelian	14?	SOV	-	?	upcoming
Hungarian	-	Uralic	18	DC	0.77	Spring 2022	42K
Indonesian	-	Austronesian	16	SVO	0.03	almost finished	168K
Japanese	-	Japonic	17	TP	0.23	?	1,680K
Mandarin Chinese	-	Sino-Tibetan	18	TP	0	Spring 2022	285K
Tamil	-	Dravidian	8?	SOV	0.97	Spring 2022	12K
Turkish	-	Turkic	18	DC	0.44	Spring 2022	591K

Table 1: CIEP language sample with details on genealogy, size of the subcorpus, word order, subject-object entropy, acquisition status and availability of Universal Dependencies parsers. Information on word order is taken from Dryer (2013) and Kiss (1995); DC stands for discourse-configurational; TP stands for topic-prominent. The entropy of subject-object order is taken from Levshina (2019). Information on Universal Dependencies parsers taken from `universaldependencies.org`.

The third issue is copy-right; CIEP+ contains almost exclusively copy-right protected materials. Under German copy-right law, we are allowed to own digital copies of these works and use them for research, but we are not allowed to share them with others. This makes dissemination of our results more difficult, but far from impossible. One of the things we are planning to do is publish mini-CIEP+, an annotated subset of corpus-material, under appropriate derived conditions. Overall, we believe creating CIEP is a valuable enterprise, and compiling and annotating it can be considered as creating a resource that is relevant and necessary in order to conduct research.

4 Acknowledgements

Thanks to the people who made CIEP+ possible in one way or another: Waleed Ahmed, Tania Avgustinova, Sahar Bahrami-Kh, Rutuja Bane, Gemma Capdevila, Teresa Cottone, Mari Dallakyan, Joanna Dietinger, Michael Dunn, Koel Dutta Chowdhury, Annika Fischer, Milen Gavrailov, Sandra Grinberga, Dmitri Hrapof, Ivan Levin, Yuchen Liao, Janani Karthikeyan, Karin Kastens, Zena Al Khalili, Ida Novindasari, Mark Pagel, Elena Parina, Eva Richter, Maryam Rajestari, Henri Ruizenaar, Algirdas Sabaliauskas, Sukanya Sengupta, Kladjola Spahiu, Meggie Uijen, Jenneke van der Wal, Ruprecht von Waldenfels, and all the people we have forgotten.

And to the funding bodies: DFG (SFB1102), ERC (268744), MPG.

References

Cappelle, Bert (Dec. 2012). "English is less rich in manner-of-motion verbs when translated from French". English. In: *Across Languages and Cultures* 13.2, pp. 173–195.

Čermák F. abd Rosen, A. (2012). "The case of InterCorp, a multilingual parallel corpus". In: *International Journal of Corpus Linguistics* 13.3, pp. 411–427.

Cysouw, Michael and Bernhard Wälchli (2007). "Parallel texts: using translational equivalents in linguistic typology". In: *STUF - Sprachtypologie und Universalienforschung* 60.2, pp. 95–99.

Dryer, Matthew S. (2013). "Order of Subject, Object and Verb". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

"Multi-CAST: Multilingual corpus of annotated spoken texts" (2021). In: ed. by Geoffrey Haig and Stefan Schnell. URL: `multicast.aspra.uni-bamberg.de/`.

Haspelmath, Martin (2018). "How Comparative Concepts and Descriptive Linguistic Categories Are Different". In: *Aspects of Linguistic Variation*. Ed. by Daniël Olmen, Tanja Mortelmans, and Frank Brisard. Berlin: De Gruyter, pp. 83–114.

Kiss, Katalin É. (1995). "Introduction". In: *Discourse configurational languages*. Ed. by Katalin É. Kiss. Oxford: Oxford University Press, pp. 3–27.

Levshina, Natalia (2016). "Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles". In: *Languages in Contrast* 16.1, pp. 84–117.

Levshina, Natalia (2019). "Token-based typology and word order entropy: A study based on Universal Dependencies". In: *Linguistic Typology* 23.3, pp. 533–572.

Marneffe, Marie-Catherine de et al. (2021). "Universal Dependencies". en. In: *Computational Linguistics* 47.2, pp. 255–308. DOI: 10.1162/coli_a_00402. (Visited on 03/03/2021).

Mayer, Thomas and Michael Cysouw (Mar. 2014). "Creating a Massively Parallel Bible Corpus". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 3158–3163. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf.

Paschen, Ludger et al. (2020). "Building a Time-Aligned Cross-Linguistic Reference Corpus from LanguageDocumentation Data (DoReCo)". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 2657–2666.

Stolz, Thomas and Traude Gugeler (2000). "Comitative Typology". In: *STUF-Sprachtypologie Und Universalienforschung* 53.1, pp. 53–61.

Tiedemann, Jörg (2012). "Parallel data, tools and interfaces in OPUS". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Verkerk, Annemarie (2014). "The evolutionary dynamics of motion event encoding". PhD thesis. Radboud University Nijmegen.

Wälchli, Bernhard (2007). "Advantages and Disadvantages of Using Parallel Texts in Typological Investigations". In: *STUF-Sprachtypologie Und Universalienforschung* 60.2, pp. 118–134.

Wälchli, Bernhard and Michael Cysouw (2012). "Lexical Typology through Similarity Semantics: Toward a Semantic Map of Motion Verbs". In: *Linguistics* 50.3, pp. 118–134. URL: https://doi.org/10.1515/ling-2012-0021.

Waldenfels, Ruprecht von (2011). "Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB". In: *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011*. Ed. by M. Daniela and R. Garabík. Bratislava, pp. 156–162.