mini-CIEP+ : A Shareable Parallel Corpus of Prose

Annemarie Verkerk and **Luigi Talamo**
Saarland University / Saarbrücken, Germany
{annemarie.verkerk, luigi.talamo}@uni-saarland.de

Abstract

In this paper we present mini-CIEP+, a sharable parallel corpus of prose. mini-CIEP+ consists of the first part of ten different works of prose across many different languages, allowing for the cross-linguistic investigation of larger discourse units. Subcorpora typically contain 5750 sentences and almost 125K tokens. Subcorpora have dependency grammar annotation based on the Universal Dependencies standard (de Marneffe et al., 2021). mini-CIEP+ version 1.0 is available in 35 languages, with the aim of increasing the sample to 50 languages. It is shareable due to recent developments in German law, which allow researchers to share up to 15% of copy-righted material with a select group of people for their own research. Hence, mini-CIEP+ is not publically available, but is rather shareable in a modular fashion with select researchers. We additionally describe future plans for further annotation of mini-CIEP+ as well as its limitations.

Keywords: parallel corpus, linguistic typology, copyright

1. Introduction

Linguistic typology, the systematic comparison of language structure across large samples of languages, has traditionally relied on discrete classifications, created by human specialists. Increasingly, however, typologists are using multilingual corpora instead: a collection of utterances (a corpus) is investigated directly using frequency-based or information theoretic measures, yielding continuous measures of language structure that are considerate of variation and sometimes, diachronic change. This approach is sometimes called token-based typology (Levshina, 2016) or corpus-based typology (Levshina, 2022; Schnell and Schiborr, 2022).

This line of work inevitably relies on the availability of cross-linguistic corpora. While many of these have emerged in the last 25 or so years (Tiedemann, 2012; Moran et al., 2022; Rosen et al., 2022; the TenTen corpora) there are distinct biases towards legalese and religious texts; and material gathered outside of those two genres often constitute (web-crawled) text fragments or collections of sentences (such as Tatoeba or the Leipzig Corpora Collection). Register is an important consideration for corpus-based typology, as we know from the study of well-described languages like English that register differences can be immense (Biber, 2012). Doing corpus-based typology solely on legal texts, web-crawled news and the Bible is at best unrepresentative of linguistic diversity.

Here we present mini-CIEP+, a sharable parallel corpus consisting of the first part of ten different works of prose. mini-CIEP+ contains subcorpora in 35 languages in version 1.0 (we aim to include 50 languages until 2028) and is annotated in the Universal Dependencies (UD) standard (de Marneffe et al., 2021). Note that while this is ongoing work, mini-CIEP+ is the first of its kind: 1) it allows for the linguistic investigation of larger discourse units (in contrast to many other web-crawled corpora); 2) the parallel nature of mini-CIEP+ has the advantage that direct comparison of subcorpora is straightforward and that annotation projection is possible (see Section 7); 3) there are no other prose corpora with this scale or size; and 4) since it contains published prose, there are no issues with variable or poor quality of the material. Given that the works of prose have copy-right, we cannot make mini-CIEP+ publicly available; however, recent changes in German law allow us to share it with other researchers. In this paper, we describe the shareable corpus as well as design and implementation choices. Corpus composition and annotation are described in Sections 3 and 4, after the overview on previous work.

2. Previous work

Since the early 2000s, several (large) parallel corpora have emerged: *EuroParl* (Koehn, 2005), *ParaSol* (Slavic prose and beyond, Waldenfels, 2006), the *Parallel Bible Corpus* (Mayer and Cysouw, 2014), *OpenSubtitles* (Lison and Tiedemann, 2016), *ParTy* (movie subtitles, Levshina, 2017), MULTEXT-East (Erjavec, 2017), *JW300* (Jehova Witness magazines, Agić and Vulić, 2019) and ParlaMint (parliamentary proceedings, Erjavec et al., 2023). Several of these have been compiled in OPUS (Tiedemann, 2012).

While these corpora contain texts from a variety of genres, most importantly legal and religious, there is a distinct lack of prose corpora, for the obvious reason that widely translated prose is typically protected under copyright law and cannot be publicly shared. Hence, the corpora used by Stolz and colleagues (for example, Stolz and Gugeler, 2000) are not publicly available and *ParaSol* (Waldenfels, 2006) can be used online but cannot be downloaded; the only exception here is *MULTEXT-East*, a parallel and morpho-syntactically annotated corpus of Orwell's *1984* in 16 languages, which is fully downloadable from the CLARIN repository[1].

Given recent changes to German and EU copyright law, some solutions for this problem have emerged. Schöch et al. (2020) propose preparing derived texts, similar in a way to datasets such as the Google Ngram Viewer[2] or the HathiTrust Research Center Extracted Features Dataset.[3] However, such datasets where

[1] http://hdl.handle.net/11356/1043
[2] https://books.google.com/ngrams/
[3] https://analytics.hathitrust.org/datasets/

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 135–143
20 May, 2024.

only frequency information or information regarding lemmas is available, but not their sequence, are not sufficient for answering many linguistic questions. Gärtner et al. (2021) propose an automated sampling approach, where users have access to 15% of individual copy-righted works (see Section 5). The downside of this approach is that samples are taken from the entirety of the text, so discourse units beyond sentences are not preserved and cannot be investigated. Bański et al. (2017) propose to make use of the long scientific quotation clause in German copyright law, arguing that a compilation of long text segments with newly created annotation constitutes a new, original work. In this case, the corpus creator enters a legal gray zone: how much annotation needs to be added in order for the corpus to be conceived as a new work?

We created mini-CIEP+ to overcome several of these problems; the legal solution is explained in Section 5. First, we describe the corpus in greater detail.

3. Corpus sample and composition

mini-CIEP+ contains a subset of the material of the Corpus of Indo-European Prose Plus (CIEP+, /kiːp plʌs/, see Talamo and Verkerk, 2022). This work-in-progress corpus will contain up to 18 literary works in 50 languages, with a bias towards Indo-European. mini-CIEP+ contains about 14% of 10 of these literary works (see Section 5) in the same languages:

1. IE, Albanian: Standard Albanian
2. IE, Armenian: Eastern Armenian
3. IE, Baltic: Latvian, Lithuanian
4. IE, Celtic: Breton, Irish, Welsh
5. IE, Germanic: Afrikaans, Danish, Dutch, English, German, Swedish
6. IE, Hellenic: Modern Greek
7. IE, Indo-Aryan: Assamese, Bengali, Hindi, Marathi, Nepali, Punjabi, Sinhala, Urdu
8. IE, Iranian: Kurdish, Persian
9. IE, Romance: French, Latin, Italian, Portuguese, Romanian, Spanish
10. IE, Slavic: Bulgarian, Czech, Polish, Russian, Serbo-Croatian, Ukrainian
11. Austronesian: Hawaiian, Indonesian, Maori
12. Bantu: Swahili
13. Basque
14. Dravidian: Tamil
15. Japonic: Japanese
16. Kartvelian: Georgian
17. Koreanic: Korean
18. Semitic: Arabic
19. Sinitic: Mandarin Chinese
20. Turkic: Turkish
21. Uralic: Finnish, Hungarian

Given that the translation of prose is driven by monetary impetuses, the mini-CIEP+ language sample is biased towards European and other well-described languages (see Wälchli, 2007). The prose works chosen have been selected first for their popularity, i.e. because they have been widely translated, and second, for being originally written in different languages, so as to avoid English as the sole source language. We are aware that the original texts are written exclusively in Indo-European languages, more specifically, in French, Italian, Spanish and Portuguese (Romance), Dutch, German and English (Germanic) and Modern Greek (Hellenic). Sadly, this bias cannot be avoided; out of the titles listed under the Wikipedia entry 'List of literary works by number of translations'[4], there are about 80 books that can be loosely classified as 'prose', namely, novels, diaries and plays; however, the majority are originally written in languages from the three above-mentioned branches, especially English. Other works of prose that could be considered come with certain difficulties. Children's stories such as *Pinocchio* often suffer from abridged translations. Books not originally written in one of the languages mentioned above are few; those that exist, such as *The Upright Revolution: Or Why Humans Walk Upright* (by Ngũgĩ wa Thiong'o), are either too short, not modern (*The tragedy of Man*, by Imre Madách), or very hard to obtain (such as Ismail Kadare's *The General of the Dead Army*).[5]

Given these considerations, mini-CIEP+ contains the first part of the following ten texts.[6] A list of authors, titles, and date of first publication is provided here for brevity; an overview of mini-CIEP+ is available in Table 1. Acronyms refer to columns in that Table.

1. **AA** – Carroll's *Alice's Adventures in Wonder-land* [English, 1865]
2. **LG** – Carroll's *Through the Looking-Glass and What Alice Found There* [English, 1871]
3. **Al** – Coelho's *O Alquimista* [The Alchemist, Portuguese, 1989]
4. **Za** – Coelho's *O Zahir* [The Zahir, Portuguese, 2005]
5. **Ro** – Eco's *Il nome della rosa* [The Name of the Rose, Italian, 1980]
6. **Di** – Anne Frank's *Het Achterhuis* [Diary of a Young Girl, Dutch, 1947][7]
7. **100Y** – García Márquez's *Cien Años de Soledad* [One Hundred Years of Solitude, Spanish, 1967]
8. **Zo** – Kazantzakis' *Βίος και Πολιτεία του Αλέξη Ζορμπά* [Zorba the Greek, Modern Greek, 1946]
9. **Pr** – de Saint-Exupery's *Le Petit Prince* [The Little Prince, French, 1943]
10. **Pa** – Süskind's *Das Parfum. Die Geschichte eines Mörders* [Perfume: The Story of a Murder-er, German, 1985]

[4] https://en.wikipedia.org/wiki/List_of_literary_works_by_number_of_translations

[5] Another concern might be how modern the corpus is, given that AAiW and TtLG are from the late nineteenth century, and we have several books from the 1940s and 1980s. However, all of these are considered modern classics and many translations we have obtained are far more recent than these first dates of publication betray.

[6] All originals are included. When selecting the translations, we aim for the most recent one or one which has been translated directly from the original (non-mediated).

[7] Of course, Anne Frank's *Het Achterhuis* is not a work of fiction. We include it because it is the most widely translated Dutch original text, and because in terms of its register, it is not far from the other included texts. Diary entries can be considered stories told from a first-person perspective.

Family, genus	Language	100Y	AA	Di	Al	Ro	Pa	Pr	LG	Za	Zo	T	UD
IE Celtic	**Welsh**	-	1	1	-	-	-	1	1	-	-	4	p
IE Celtic	**Irish**	-	1	-	1	-	-	1	1	-	-	4	p
IE Indo-Aryan	**Urdu**	-	-	-	1	-	-	1	-	1	1	4	p
IE Romance	**Latin**	-	1	-	-	-	1	1	1	-	-	4	p
IE Germanic	Afrikaans	-	1	1	1	-	-	1	-	-	1	5	p
Dravidian	Tamil	1	1	1	1	-	-	1	-	1	-	6	p
IE Indo-Aryan	**Marathi**	1	1	1	1	-	-	1	-	1	-	6	p
Basque	**Basque**	-	1	1	1	-	1	1	1	-	1	7	p
IE Armenian	**Armenian**	1	1	1	1	1	1	1	1	-	1	9	p
IE Indo-Aryan	**Hindi**	1	1	1	1	1	-	1	1	1	1	9	p
Austronesian	**Indonesian**	1	1	1	1	1	1	1	1	1	-	9	p
IE Hellenic	**Modern Greek**	1	1	1	1	1	1	1	1	1	1	10	p
IE Baltic	**Latvian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Baltic	**Lithuanian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	**Swedish**	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	**Danish**	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	**Dutch**	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	**English**	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	**German**	1	1	1	1	1	1	1	1	1	1	10	p
IE Iranian	**Persian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	**Portuguese**	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	**French**	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	**Italian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	**Romanian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	**Spanish**	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	**Bulgarian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	**Serbo-Croatian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	**Czech**	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	**Polish**	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	**Russian**	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	**Ukrainian**	1	1	1	1	1	1	1	1	1	1	10	p
Uralic	**Finnish**	1	1	1	1	1	1	1	1	1	1	10	p
Uralic	**Hungarian**	1	1	1	1	1	1	1	1	1	1	10	p
Japonic	Japanese	1	1	1	1	1	1	1	1	1	1	10	p
Semitic	**Arabic**	1	1	1	1	1	1	1	1	1	1	10	p
Sinitic	**Man. Chinese**	1	1	1	1	1	1	1	1	1	1	10	p
Turkic	**Turkish**	1	1	1	1	1	1	1	1	1	1	10	p
Koreanic	Korean	1	1	1	1	1	1	1	1	1	1	10	p
IE Celtic	Breton	-	1	-	-	-	-	1	-	-	-	2	t
IE Indo-Aryan	Assamese	-	1	-	1	-	-	-	1	-	-	3	n
IE Indo-Aryan	Nepali	-	1	-	1	-	-	1	-	-	-	3	n
Austronesian	Maori	-	1	1	1	-	-	-	-	-	-	3	n
Austronesian	Hawaiian	-	1	-	-	-	-	1	1	-	-	3	n
Bantu	Swahili	-	1	-	1	-	-	1	-	-	-	3	n
IE Iranian	Kurdish	1	1	-	1	-	-	1	-	-	-	4	t
IE Indo-Aryan	Sinhala	-	-	1	1	-	1	1	-	1	-	5	t
IE Indo-Aryan	Bengali	1	1	1	1	-	-	1	-	1	-	6	n
IE Indo-Aryan	Punjabi	1	-	1	1	-	-	1	-	1	1	6	n
IE Albanian	**Albanian**	1	1	1	1	1	1	1	1	1	1	10	t
Kartvelian	Georgian	1	1	1	1	1	1	1	1	1	1	10	n

Table 1: Overview of literary works available per language in mini-CIEP+. The last column, "UD", specifies relevant information regarding UD (Universal Dependencies) version 2.13: p (pre-trained model available in Stanza (Qi et al., 2021)), t (treebank available without pre-trained model) and n (no UD treebank available). The languages printed in bold are included in mini-CIEP+ version 1.0.

If all ten books are available, the size of the subcorpus for a single language is approximately 121,000 tokens, or 5750 sentences. The size of each subcorpus is provided in Table 2, in terms of both tokens and sentences – two statistics on key UD dependency labels are also given. However, note that not all ten books are available in all fifty languages (see Table 1). Most or all works of prose are available in most languages, but for some languages only four or fewer are available. In order to have approximately equal subcorpora sizes, we add more prose works to a subcorpus such as that of Irish, which only contains four out of the ten prose works listed above.[8] Hence, with the addition of two translated works and four native Irish works, the Irish subcorpus has become a comparable rather than a parallel subcorpus – in the sense that the added texts are translated and original prose. In these cases, we aim to obtain at least the English translations or originals, so the paired subcorpora can be used for contrastive analyses.

4. Corpus processing and annotation

The CIEP+ corpus exists both physically and digitally. The first step to obtain the relevant textual material for each subcorpus is to obtain a physical copy of each book (see Section 5) and create or buy in addition a digital version. In most cases, the digital version is created by scanning the book and applying Optical Character Recognition (OCR) to retrieve the contents in plain text format. The result has to be checked and corrected by human annotators, as automated OCR usually generates a lot of mistakes.

Then, the texts that are included in each subcorpus are annotated with metadata for the following information: original author, original title, original publishing date, original language, translator, translation language, translation title, translation date and translation publishing house. The physical books are cataloged in the university library (SULB).

It is not feasible to provide morphosyntactic annotation of such a large and diverse data set by hand. Hence, the first layers of annotation are added automatically. We have chosen to do this within the Universal Dependencies (UD) framework (de Marneffe et al., 2021), for several reasons. Firstly, UD's aim of providing consistent annotation of morphosyntax (including parts of speech, morphology, and syntactic dependencies) across different languages aligns with our own: we need consistent morphosyntactic annotation in order to use the data to ultimately answer typological research questions. The Universal Dependencies project is emerging as the go-to set of treebanks for typologists, given its wide sample of parsed language data, which we (and others) use not only for doing typology on, but also for training tools that can automatically parse new language data. Secondly, dependency grammar is central to our goals in the larger project, given that we are interested in dependency length optimization and other functional metrics of language-in-use (see Dyer, 2023). Thirdly, given the status of UD as emerging standard of the field implies that there exist a lot of (also future) resources that allow us to parse additional languages (see below), but that also allow prospective users of mini-CIEP+ to convert it to formats of their choice.

The tool chosen to process corrected texts and create automated annotation in the UD standard is the Stanford Stanza natural language analysis package[9] (Qi et al., 2021). Among the 24 systems participating in the CoNLL 2018 Shared Task (Zeman et al., 2018), Stanford Stanza ranked eight in the labeled attachment score (LAS), second in the Morphology-Aware Labeled Attachment Score (MLAS) and fifth in the Bilexical Dependency Score (BLEX); to the best of our knowledge, only two systems that performed slightly better than Stanza are currently available to the community, UDPipe Future[10] (Straka, 2018; now UDPipe 2) and Turku NLP[11] (Kanerva et al., 2018). However, in the CoNLL 2018 Shared Task systems were evaluated on Universal Dependencies treebanks, which widely differ from mini-CIEP+ data in terms of register. We leave for future work a shared task performed on mini-CIEP+ data, comparing Stanford Stanza to other available systems.

At the time of writing, Stanza comes with 138 models, which are pretrained on Universal Dependencies version 2.13 treebanks and cover 38 languages of the sample. These models are used to parse the corrected texts, processing and annotating them in several steps, including sentence splitting, tokenization, lemmatization, parts-of-speech and syntactic dependencies tagging, and, where available, multi-word token expansion and named entity recognition.

This leaves twelve languages without pre-trained Stanza models (see also Table 1). As for some of these low resource languages, we have used small existing Universal Dependencies treebanks to train parsers for three languages, namely for Breton, Kurdish, and Sinhala (results are not included in mini-CIEP+ v. 1.0, but will be in later versions). While we have not formally evaluated these so far, results very much depend on the size (and register) of the UD treebank.

This leaves nine languages in our sample with no or highly limited Universal Dependencies resources (see Table 1). We ourselves started projects to provide resources for two of the low resource languages –

[8] For Irish, we have added six texts in order to try to approach a similar token size as the other subcorpora:
1. *An Béal Bocht* (The Poor Mouth), Flann O'Brien
2. *An Hobad, nó Anonn agus Ar Ais Arís* (The Hobbit, or There and Back Again), J. R. R. Tolkien
3. *An Leon, an Bandraoi agus an Prios Éadaigh* (The Lion, the Witch and the Wardrobe), C. S. Lewis
4. *Buille Marfach* (A Fatal Blow), Anna Heussaff
5. *Cré na Cille* (Graveyard Clay), Máirtín Ó Cadhain
6. *Rún an Bhonnáin* (The secret of the Bonnán), Proinsias Mac a' Bhaird

[9] https://stanfordnlp.github.io/stanza/
[10] https://github.com/ufal/udpipe/releases/tag/v2.1.0
[11] https://turkunlp.org/Turku-neural-parser-pipeline/

Language	Bks	Token	Sent.	nsubj	obj		Language	Bks	Token	Sent.	nsubj	obj
Albanian	10	135158	6401	8493	10659		Latin	3	9003	670	718	610
Arabic	10	123994	NA	8649	5689		Latvian	10	105635	6234	10023	7506
Armenian	6	68696	3503	4785	4030		Lithuanian	10	105226	6800	7964	4287
Basque	4	19870	1244	1013	1461		Man. Chinese	10	136777	6064	13038	9824
Bulgarian	10	118040	6369	6997	9742		Marathi	8	105197	5990	9208	7629
Czech	10	114149	6263	6868	5555		Persian	10	131749	6039	7058	5316
Danish	10	133082	6250	13478	8772		Polish	10	116429	6228	6011	8820
Dutch	10	133933	6243	12584	6710		Portuguese	10	135648	6281	6903	8400
English	10	138386	6472	12802	6794		Romanian	10	131484	5668	7051	7862
Finnish	5	43335	3278	4067	2529		Russian	10	117115	6245	9868	5868
French	10	144199	6365	11904	9267		Serb.-Croatian	10	115582	5888	7270	7475
German	10	130730	6139	12232	7308		Spanish	10	130947	5731	6113	7633
Mod. Greek	10	125972	5393	6601	6132		Swedish	8	97054	4468	10299	5670
Hindi	8	95667	4536	9147	5532		Turkish	10	94958	5633	6647	7214
Hungarian	10	110675	5812	7378	6804		Ukrainian	10	109248	5757	9085	6721
Indonesian	9	104799	5089	9694	6581		Urdu	4	38217	2368	3456	2243
Irish	10	56920	3165	4896	1981		Welsh	4	33611	1494	2333	1267
Italian	10	137672	6168	6485	7379							

Table 2: Overview of descriptive statistics of mini-CIEP+ version 1.0. Bks = Books; Token = Tokens; Sent = Sentences. nsubj and obj refer to the number of constituents with these labels in each parsed subcorpus. Some subcorpora still lack some texts that have to be processed (see Section 4), which will be part of mini-CIEP+ version 1.1. Further languages listed in Section 3 and Table 1 will be included in future versions.

these projects take the form of manually annotated UD treebanks covering literary works originally written in Albanian (Talamo, in prep.) and Bengali (Dyer, in prep. b). These treebanks are used to train good quality parsers, specifically aimed to the genre featured in our parallel corpus, and allow for automated parsing of the Albanian and Bengali subcorpora. Although others are similarly spearheading solutions for the lack of resources in several languages, this will remain problematic in years to come. This means that seven languages of our sample (Assamese, Georgian, Hawaiian, Maori, Nepali, Punjabi, Swahili) do not currently have any existing UD treebanks; for these languages, we wait for relevant UD treebanks to become available, or find alternative solutions. Such solutions will include zero-shot analysis alongside corrections, for example using UDify (Kondratyuk and Straka, 2019), and converting existing treebanks to the UD standard.

UD's native CoNLL-U format allows for additional annotation in the last column, and the newer CoNLL-U Plus format allows for even more columns. We aim to release versions of mini-CIEP+ with surprisal and information status annotation (see Section 7). For users of mini-CIEP+, these columns can be used for other types of annotation. The modular nature of the corpus also allows for re-parsing with better models and human correction of automated annotation.

5. Sharing the corpus

Given its size and its cost in terms of resources, we did not wish to create CIEP+ (the Corpus of Indo-European Prose Plus) only for project internal purposes (see also Hartmann's 2023 proposal on Open Corpus Linguistics). German copyright law has changed in 2018 regarding two important aspects: collecting copyrighted material for research and sharing it with a select group of people. The relevant articles are Urheberrecht § 60c and 60d.[12] Under German law, we are allowed to store digital copies of copy-righted works and use these for research if we own the physical books. Then, most relevant for mini-CIEP+ is the following sentence; original German in the footnote below:

"*For the purpose of non-commercial scientific research, up to 15 percent of a work may be reproduced, distributed and made publicly accessible [...] to a defined circle of people for their own scientific research*"[13]

[12] https://www.gesetze-im-internet.de/urhg/__60c.html https://www.gesetze-im-internet.de/urhg/__60d.html

[13] "Zum Zweck der nicht kommerziellen wissenschaftlichen Forschung dürfen bis zu 15 Prozent eines Werkes vervielfältigt, verbreitet und öffentlich zugänglich gemacht werden

Hence, we created mini-CIEP+ to legally share 15% of CIEP+ with a specifically designated group of people in order to benefit their research. We believe that there is indeed a pluricentric group of people that would benefit from mini-CIEP+: corpus-based typologists, but also contrastive linguists and language specialists. As we include several low-resource languages, it is our hope that parts of mini-CIEP+ can be used for furthering research into those languages.

To make this possible we have created a data usage agreement (see Appendix A) that specifies the conditions under which mini-CIEP+ can be provided to potential data users. This data usage agreement also asks which subcorpora are needed by the researcher, so that the corpus is really only shared to the extent required. Version management takes place via the author's home page,[14] so that prospective data users know what is available for sharing.

6. CIEP+-based works so far

CIEP+ (the Corpus of Indo-European Prose Plus) is being built in the context of a large research program,[15] in which our team have authored half a dozen of papers in the last four years. In this section, we give an overview of these papers in order to showcase what type of linguistic research can be done on such a resource. As primarily a resource for typologists, CIEP+ was first exploited for addressing one of the oldest topics in linguistic typology, namely, word order variation. Talamo and Verkerk (2022) investigated the order of constituents in five nominal constructions (the order of article, demonstrative, adjective, adposition and relative clause with respect to the noun) in a sample of 11 Indo-European languages, using Shannon's entropy as a metric for word order variability. The results show the high unpredictability of the position of adjectives in Romance and Slavic languages, while the entropy of constructions like determiners and adpositions is generally low. The latter confirms the traditional view of categorical studies; however, there are in fact outliers, as we retrieve phenomena that create variability in the position of prepositions in Dutch and find a certain degree of freedom for demonstratives in Greek, Polish and Welsh.

Talamo (2023) has further expanded the research regarding word order variability within the noun phrase by looking at neglected and hard-to-catch categories such as quantifiers, determiners and numerals; in a sample of 17 languages, Talamo (2023) finds that the variability of demonstratives is found in another Balkan language, Romanian, and reports on the high variability of quantifiers in Irish.

In the field of historical linguistics, Talamo et al. (2024) used CIEP+ to challenge the traditional view which states that subordinate clauses tend to preserve more conservative features than main clauses. Focusing on adverbial clauses and using frequency data on null subject pronouns and order of subject, object and verb in a sample of 30 Indo-European languages, they show that there are actually very few asymmetries between adverbial and main clauses, both in the synchronic data and during language change, which is modelled using phylogenetic methods.

The prose genre of CIEP+, which is characterized by several dialogues mimicking the spoken language, allows for research into linguistic devices used for reference. In an ongoing study (Steuer et al. in prep.), we are exploring the relations between personal pronouns and their referents, trying to understand how the former encode the information status of the latter. We model the probability in context (surprisal) of personal pronouns in a sample of 15 languages from eight different families using mGPT (Shiliazkho et al. 2022). We expect that these models reflect varying surprisal of personal pronouns based on their frequency and usage patterns, showing that first and second personal pronouns encode less information than third personal pronouns.

Several of these studies, including Talamo et al. (2024) and Levshina et al. (2023), contain comparisons between CIEP+ and UD treebanks. We can confirm that automatically parsed data from CIEP+ behaves similarly (i.e. is correlated with) data from Universal Dependencies treebanks on several measures, including word order variation and pronoun usage. However, there are notable differences between the two data sources, especially concerning individual languages on certain measures. We leave for future work a systematic comparison of CIEP+ and mini-CIEP+ with UD treebanks, with the specific aim of investigating if such differences are rooted in register differences, problems with automated parsing, or inconsistencies in UD annotation across languages.

7. Future plans

Currently, mini-CIEP+ is automatically annotated using the UD framework (de Marneffe et al., 2021, see above) in the same way as CIEP+. However, as mentioned above, we aim to add several types of annotation to mini-CIEP+, which can be shared in future versions. One type of annotation that we aim to add to CIEP+ and mini-CIEP+ is sentence and word alignment. This is obviously a great asset for a parallel corpus, however, performance on automated alignment will vary radically from language pair to language pair. While the pivot language will be English, we will carry out experiments to see if automated sentence alignment can be improved by employing different or even multiple pivots. Alignment is necessary in order to be able to project different types of annotation across the subcorpora. We will focus on information status annotation. Ongoing work (Dyer in prep. a) is preparing information status

[...] für einen bestimmt abgegrenzten Kreis von Personen für deren eigene wissenschaftliche Forschung"

[14] https://www.uni-saarland.de/lehrstuhl/verkerk.html

[15] https://sfb1102.uni-saarland.de

annotation using human annotators for English, modern Greek, Indonesian, Turkish, and Ukrainian. This version of mini-CIEP+ can also be shared with researchers interested in such annotation.

If data users require us to do so, it is possible to add more languages to the sample, especially for *Alice's Adventures in Wonderland* and *Le Petit Prince*, as these are the corpus' most widely translated books.

8. Conclusion and limitations

We have presented mini-CIEP+, a sharable parallel corpus of prose. We have described its compilation, composition, size, annotation, and plans on how to share it with relevant researchers. This is the first version and more versions are planned for the future.

We conceive of mini-CIEP+ as a modular resource for corpus-based typologists, contrastive linguists and language specialists. Individual subcorpora may not be large (~121,000 tokens), but they are large enough to research a plethora of linguistic phenomena, including semantic and pragmatic features that emerge only in the analysis of bigger discourse units. We hope that mini-CIEP+ will be used and expanded, if so, we will do our best to expand it further in a way that benefits the scientific community. Including other books for individual subcorpora would be possible.

One limitation we cannot fix is the inherent bias in the sample of languages. mini-CIEP+ is a derivative of CIEP+ (the Corpus of Indo-European Prose Plus); the inclusion of mostly Indo-European languages is intentional but at the same time, a regrettable continuation from similar biases in other corpora. Aside from *Le Petit Prince* and, to a lesser extent, *Alice's Adventures in Wonderland,* the corpus' set of prose texts (indeed, published prose in general) tends to be translated in only a very small subset of the world's languages. A positive outlook on this is offered by the larger amount of variety included in the Universal Dependencies treebanks, and in other projects such as TeDDi (Moran et al., 2022). A worthwhile solution is for corpus-based typologists to find ways to be able to analyze heterogeneous data sources, possibly with the help of NLP tools. These will not always have the same register, annotation, size, or even script, but combining (still scarce) resources on the languages of the world will be essential in future ventures in quantitative typology.[16]

9. Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

10. Bibliographical References

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Florence. Association for Computational Linguistics.

Bański, P., Kamocki, P., and Trawiński, S. (2017). Legal canvas for a patchwork of multilingual quotations: the case of CoMParS. Presented at the Corpus Linguistics International Conference 2017, Birmingham.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1): 9–37.

Dyer, A. (2023). Revisiting dependency length and intervener complexity minimisation on a parallel corpus in 35 languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 110–119.

Dyer, A. (in preparation a). CiepInf: A multilingual parallel corpus for coreference resolution and information status in the literary domain. [working title]

Dyer, A. (in preparation b). The Shobdokosh dependency corpus of Bengali prose. [working title]

Erjavec, T. (2017). MULTEXT-East. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation*. Dordrecht: Springer, pp. 441-462.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Calzada Pérez, M., de Macedo, L.D. Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darǵis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation* 57(1): 415-448.

Gärtner, M., Kleinkopf, F., Andresen, M., and Kupietz, M. (2021). Corpus reusability and copyright – challenges and opportunities. In H. Lüngen, M. Kupietz, P. Bański, A. Barbaresi, S. Clematide, & I. Pisetta (Eds.), *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021*, pages 10–19. Leibniz-Institut für Deutsche Sprache.

Hartmann, S. (2023). Open corpus linguistics – or how to overcome common problems in dealing with corpus data by adopting open research practices. Preprint. PsyArXiv.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In D. Zeman & J. Hajič (Eds.) *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133-142. Association for Computational Linguistics.

[16] Author contributions. AV: conceptualization; validation; data collecting; writing (original draft) sec- tions: 1, 2, 3, 4, 6, 7, 8, Appendix; writing (review & editing); LT: validation; data collecting; data parsing; writing (original draft) sections: 3, 4, 5; writing (review & editing).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers*, pages 79-86. Phuket, Thailand.

Kondratyuk, D., and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally'. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779-2795. Hong Kong. Association for Computational Linguistics.

Levshina, N. (2016). Why we need a token-based typology: a case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2): 507–542.

Levshina, N. (2017). A multivariate study of T/V forms in European languages based on a parallel corpus of film subtitles. *Research in Language* 15(2): 153–172.

Levshina, N. (2022). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(2): 129-160.

Levshina, N., Namboodiripad, S., Allassonnière-Tang, M., Kramer, M.A., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gupton, T., Kidd, E., Liu, Z., Naccarato, C., Nordlinger, R., Panova, A., and Stoynova, N. (2023). Why we need a gradient approach to word order. *Linguistics*, 61(4): 825-883.

Lison, P., and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).

de Marneffe, M.-C., Manning, C.D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics* 47(2): 255–308.

Mayer, T., and Cysouw, M. (2014). Creating a massively parallel Bible corpus. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik. European Language Resources Association (ELRA).

Moran, S., Bentz, C., Gutierrez-Vasques, X., Sozinova, O., and Samardzic, T. (2022). 'TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1150–1158. Marseille. European Language Resources Association (ELRA).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C.D. (2020).'Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Online. Association for Computational Linguistics.

Rosen, A., Vavřín, M., and Zasina, A.J. (2022) *InterCorp*, Release 15 of 11 November 2022. Institute of the Czech National Corpus, Charles University. Available from: http://www.korpus.cz.

Schnell, S. and Schiborr, N.N. (2022). Crosslinguistic corpus studies in linguistic typology'. *Annual Review of Linguistics* 8(1): 171–191.

Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., and Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geistes-wissenschaften*.

Shliazhko, O., Fenogenova, A., Tikhonova, M., Mikhailov, V., Kozlova, A., and Shavrina, T. (2022). mGPT: Few-shot learners go multilingual. *arXiv* preprint arXiv:2204.07580.

Steuer, J., Talamo, L., and Verkerk, A. (in preparation). Measuring accessibility through surprisal: a cross-linguistic study of personal pronouns. [working title]

Stolz, T., and Giugeler, T. (2000). Comitative typology – nothing about the ape, but something about king-size samples, the European community and the little prince. *STUF - Sprachtypologie und Universalien-forschung* 53(1): 53–61.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197-207. Brussels. Association for Computational Linguistics.

Talamo, L. (2023). Using a parallel corpus to study patterns of word order variation: Determiners and quantifiers within the noun phrase in European languages. *Linguistic Typology at the Crossroads* 3(2): 100-131.

Talamo, L. (in preparation). STAF: The Saarbrücken Treebank of Albanian Fiction. [working title]

Talamo, L. and Verkerk, A. (2022). A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics*, 34(2), 171–226.

Talamo, L., Verkerk, A., and Salaberri, I. (2024). A quantitative approach to clause type and syntactic change in two Indo-European corpora. *Italian Journal of Linguistics*, 36.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214-2218. Istannbul. European Language Resources Association (ELRA).

von Waldenfels, Ruprecht. (2006). Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment'. In B. Brehmer, V. Zdanova, and R. Zimny (Eds.) *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*. München: Otto Sagner, pp. 123–138

Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *STUF - Sprachtypologie und Universalienforschung* 60(2): 118–134.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018) CoNLL 2018 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics*, pages 1-21. Brussels. Association for Computational Linguistics.

Appendix A. Draft of the DATA USAGE AGREEMENT FOR THE SHAREABLE PORTION OF THE CORPUS OF INDO-EUROPEAN PROSE PLUS (mini-CIEP+)

mini-CIEP+ is provided by the Corpus Provider, see below, to the Data User, as signed below, under the following terms:

1. mini-CIEP+ may only be used for non-commercial linguistic research or education.

2. Usage of mini-CIEP+ is granted to individual Data Users only. All prospective Data Users of mini-CIEP+ must fill out this data usage agreement individually.

3. The Data User agrees that they will not attempt to use mini-CIEP+ to infringe on the rights of the original copyright holders; i.e. the authors/publishers of the literary works that are part of mini-CIEP+.

4. The Data User certifies that their copy of mini-CIEP+ is stored only in a single copy on computers under administration of the Data User. Data User certifies that they will take proper action for protecting this copy from being accessed, read or copied by any non-authorized person.

5. The Data User agrees to delete mini-CIEP+ after twelve months signing this agreement. The Corpus Provider must be informed that deletion of the corpus by the Data User has been done. An extension of data usage is possible by signing this agreement again.

6. mini-CIEP+ is provided free of charge.

7. mini-CIEP+ comes with absolutely no warranties including (but not limited to) the correctness of the information provided in the text corpus itself.

8. The Data User will not disclose, disseminate, or otherwise share mini-CIEP+ to or with any other person or entity, for any purpose. The Data User has no right to copy, redistribute, transmit, publish or otherwise use mini-CIEP+ for any other purpose.

9. mini-CIEP+ must not be transmitted electronically to other services not under administration of the Data User, such as online translation services.

10. The Data User may include limited excerpts from mini-CIEP+ in articles, reports and other documents describing the results of the Data User's non-commercial linguistic education or research.

11. In no event shall the Corpus Provider be liable to the Data User for direct, indirect, special, incidental, punitive or consequential damages of any kind arising in any way out of this agreement, rights granted herein or by the use of mini-CIEP+.

12. mini-CIEP+, in all forms, shall be and remain the responsibility of the Corpus Provider.

13. The Data User will provide the Corpus Provider with a short summary (less than 100 words, see below) describing the purpose of their research based on mini-CIEP+ and the language sample they require. The Data User agrees that all their actual research activities with mini-CIEP+ will adhere to this description. Using mini-CIEP+ for a different kind of research requires signing a new data usage agreement with a new description.

14. The Data User agrees that their name, contact information, and the research summary are stored in electronic form by the Corpus Provider. This information will be used to (a) inform Data Users when updates of mini-CIEP+ are available and to (b) create anonymized corpus distribution statistics. Additionally, the information might be used to track violations of this agreement. It will be deleted once this Data Usage Agreement is expired or cancelled by the Data User or by the Corpus Provider.

15. Contributions which are based on mini-CIEP+ must cite the following publication: <xxx>

16. Contributions which are based on mini-CIEP+ must correctly cite its version as well as the original works compiled in mini-CIEP+, which can be retrieved from mini-CIEP+'s metadata.

17. The Data User shall email an electronic version of the signed agreement to the Corpus Provider.