Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 805–813
Marseille, 20-25 June 2022

ParCorFull2.0: a Parallel Corpus Annotated with Full Coreference

Ekaterina Lapshinova-Koltunski[1], Pedro Ferreira[2], Elina Lartaud[3], Christian Hardmeier[3,4]
[1]Saarland University, [2]University of Aveiro, [3]Uppsala University, [4]IT University of Copenhagen
e.lapshinova@mx.uni-saarland.de, pedroaferreira@ua.pt, elina.aurelia@gmail.com, chrha@itu.dk

Abstract

In this paper, we describe ParCorFull2.0, a parallel corpus annotated with full coreference chains for multiple languages, which is an extension of the existing corpus ParCorFull (Lapshinova-Koltunski et al., 2018). Similar to the previous version, this corpus has been created to address translation of coreference across languages, a phenomenon still challenging for machine translation (MT) and other multilingual natural language processing (NLP) applications. The current version of the corpus that we present here contains not only parallel texts for the language pair English-German, but also for English-French and English-Portuguese, which are all major European languages. The new language pairs belong to the Romance languages. The addition of a new language group creates a need of extension not only in terms of texts added, but also in terms of the annotation guidelines. Both French and Portuguese contain structures not found in English and German. Moreover, Portuguese is a pro-drop language bringing even more systemic differences in the realisation of coreference into our cross-lingual resources. These differences cause problems for multilingual coreference resolution and machine translation. Our parallel corpus with full annotation of coreference will be a valuable resource with a variety of uses not only for NLP applications, but also for contrastive linguists and researchers in translation studies.

Keywords: coreference, entity reference, event reference, cross-lingual coreference resolution, coreference annotation, linguistic annotation, machine translation, multilingual NLP, multilinguality, English, German, French, Portuguese

1. Introduction

We present ParCorFull2.0[1], an extension of the existing corpus ParCorFull (Lapshinova-Koltunski et al., 2018). ParCorFull is a multilingual parallel corpus, originally containing the languages English, the original language of the included texts, and German. In this work, we describe its extension to French and Portuguese. ParCorFull has full coreference annotation, which means that it contains not only annotation of pronouns, but also full nominal phrases, verbal phrases and clauses and includes rich set of links with both entity and event coreference. The corpus was created to study, model and evaluate the translation of coreference and coherence patterns in machine translation and multilingual NLP. The first version of the corpus was already used in a number of studies (Lapshinova-Koltunski et al., 2020; Lapshinova-Koltunski et al., 2019; Guillou et al., 2018).

French and Portuguese, the two Romance languages added in the new version of the corpus, pose new annotation challenges as they contain structures not available in English and German. Therefore, the extension of the corpus with translations into these two languages also required the revision and extension of the annotation guidelines, particularly with respect to the handling of clitics and personal pronouns in direct object function, as well as relative pronouns and reflexives. Moreover, Portuguese is a pro-drop language, which means that it contains zero anaphors in subject position recognisable in the verb forms only. These language-specific features result in systemic differences in the realisation of coreference in our cross-lingual resources. Our previous analyses (Lapshinova-Koltunski et al., 2020) show that even such closely related languages as English and German show contrasts in the realisation of coreference structures across spoken and written texts. The addition of two new target languages will allow us to study more systematically how coreference is expressed and coherence is rendered across different language pairs.

The coreference relation is shared across all languages, but languages differ considerably in the range of linguistic means triggering this relation (Lapshinova-Koltunski et al., 2020; Lapshinova-Koltunski and Kunz, 2020; Kunz and Steiner, 2012; Kunz and Lapshinova-Koltunski, 2015; Novák and Nedoluzhko, 2015). The more differences there are in the language systems, the more variation in coreference means we observe, even if we deal with the same pieces of information.

(1) a. EN: *...not to mention social networking platforms, allow [people] to self-identify, to claim [their] own descriptions of [themselves], so [they] can go align with global groups of [their] own choosing.*

b. DE: *...gar nicht zu sprechen von Social Networking Plattformen, [Menschen] ermöglichen [sich] eine eigene Identität geben, [sich] auf eigene Art und Weise definieren, und sich damit weltweit an zu Gruppen orientieren, die [sie] [sich] selbst aussuchen.*

[1] The corpus will be available from the LINDAT repository. The data is already available at the GitHub repository `https://github.com/chardmeier/parcor-full`.

c. PT: *...já para não falar nas plataformas sociais na internet, permitem [às pessoas] [auto-identificarem-se], e criarem as descrições de [si próprias] de maneira a [lhes] permitir [associarem-se] globalmente aos grupos que [quiserem].*

d. FR: *...sans parler des plateformes de réseaux sociaux, permettent [aux gens] de s'identifier [eux]-mêmes, de revendiquer leur propre description d'[eux]-mêmes, de manière à pouvoir rejoindre les groupes mondiaux de [leur] choix.*

Example (1) illustrates the same coreference chain in English (EN), German (DE), Portuguese (PT) and French (FR). All expressions refer to the entity *people* (*Menschen, pessoas, gens*). In the English sentence (1-a), the chain contains a bare noun as antecedent and two possessive pronouns, one personal and one reflexive pronoun are anaphors. Its German (1-b) translation also contains a bare noun as antecedent. However, the German chain does not contain any possessive pronouns – reflexive pronouns are used instead. In the Portuguese translation (1-c), we find more variation: The antecedent *às pessoas* is a definite noun phrase with a preposition fused to the definite article. It contains three reflexives (which are parts of the verbs here), one personal pronoun and one pro-drop or zero anaphora (also marked on the verb). French (1-d), like Portuguese, has a definite NP antecedent with a fused preposition Moreover, the finite subordinate clause *so [they] can go align* at the end of the example is rendered with an infinite construction *de manière à pouvoir rejoindre*, causing the disappearance of the subject pronoun in the translation.

Such cross-lingual differences in the realisation of coreference relation give rise to transformation patterns used to create coherent translations. Therefore, understanding them is valuable for not only contrastive linguistics and translations studies, but also for multilingual natural language processing.

The remainder of the paper is organised as follows. In Section 2., we summarise the existing corpora annotated with coreference for the languages at hand. Section 3. provides an overview of the categories annotated and contains information on the new categories that we introduced to capture specific properties of the two Romance languages. We give some details on the selection of the French and Portuguese data in Section 4. and on the annotation process in Section 5. Section 6. contains statistics on the annotated structures. An outlook and conclusions are given in Section 7.

2. Related Work

There exist several corpora annotated with coreference relations for the English language. For German, the resources are more sparse. There are a few corpora for French and Portuguese. However, we only know of few multilingual corpora for multiple languages that are annotated with the same coreference categories. We summarise some of the monolingual and multingual corpora known to us below.

English corpora The corpus ARRAU (Uryupina et al., 2020) is one of the most thoroughly coreference-annotated English corpora. It contains large-scale annotations of a wide range of anaphoric phenomena in texts belonging to various genres including news, dialogues and fiction.thorough annotation.

Another corpus containing various genres is GUM (Zeldes, 2017). The coreference annotations in GUM additionally provide details on the structural information status (given, accessible, new) of mentions. All named entities (and the further mentions) are also linked to their Wikipedia identifier provided they have a Wikipedia article.

The corpus OntoNotes (Weischedel et al., 2017) is one of the most well-known large-scale resources annotated with coreference. This resource is multilingual and contains English, Chinese and Arabic texts from news, magazines, web data, broadcast conversations and conversational speech data.

The English part of PCEDT (Nedoluzhko et al., 2016), the Prague Czech-English Dependency Treebank, contains the Wall Street Journal section of the Penn Treebank and includes coreference annotations produced in the same way as those in the Czech-PCEDT.

The corpus TwiConv (Aktaş and Kohnert, 2020) contains of coreference annotation of microblog conversations from Twitter.

The Parallel Meaning Bank (Abzianidze et al., 2017), a semantically annotated parallel corpus for English, German, Dutch and Italian, also contains coreference annotations.

German corpora As already mentioned above, there are not so many coreference-annotated corpora for the German language.

Coreference annotations are contained in TüBa/DZ (Naumann and Möller, 2007). PotsdamCC (Bourgonje and Stede, 2020), a corpus of German newspaper articles, is annotated with nominal and pronominal coreference. The relations are restricted to the category of identity only. Besides that, the corpus contains information on the information structure.

There are several multilingual corpora that also contain coreference annotations of German texts. Apart from the first version of ParCorFull, there is ParCor (Guillou et al., 2014) with a pairwise annotation of anaphoric pronouns and their antecedents, a very small corpus by (Grishina and Stede, 2015) and the corpus GECCo (Kunz et al., 2021). These four corpora contain English and German texts, with German texts being translations of the English sources in most cases. The corpus GECCo contains comparable texts in English and German.

French corpora The corpus DeDe (Gardent and Manuelian, 2005) was one of the first freely available coreference-annotated corpora for French. This corpus contains newspaper articles.

Another freely available corpus for French that contain coreference annotations is ANCOR (Muzerelle et al., 2014). This corpus represents the largest French corpus that concerns specifically spoken language, as it contains a variety of spoken genres. The coreference annotations include nominal and pronominal mentions.

Democrat (Landragin, 2016) is a multi-genre corpus of written texts. This corpus partly contains historical data. Both ANCOR and Democrat were specifically designed to tackle the coreference resolution task.

EvalRefGen (Todirascu, without year) is a small multi-genre corpus of about 15,000 tokens annotated for primarily nominal coreference.

The ELRA-W0032 corpus (Tutin et al., 2000) contains one million tokens of mostly journalistic text and two monographs annotated for anaphoric and deictic expressions as well as some cases of ellipsis and event reference.

Portuguese corpora There exist several corpora annotated with coreference in Portuguese.

The HAREM corpus (da Fonseca et al., 2017) contains annotations of nominal coreference only, as it was created for automatic detection of relations between named entities. This was one of the first joint evaluation efforts for Portuguese.

The Summ-it corpus (Collovini et al., 2007) contains not only annotations of coreference relations, but also information on morpho-syntactic properties of referring expressions. There are 560 coreference chains marked in this corpus.

Corref-PT (Vieira et al., 2018) is another Portuguese corpus annotated with coreference, which contains 3,898 reference chains. They were automatically annotated and then manually revised. However, all these corpora are monolingual.

To our knowledge, the only multilingual corpus containing annotation of coreference for PT is described in (Garcia and Gamallo, 2014). This corpus includes Portuguese, Galician and Spanish. However, coreference annotations are restricted to person entities only.

The ZAC (Zero Anaphora Corpus) is a corpus compiled with the aim to resolve zero-anaphora, that is, an anaphora relation where the anaphoric expression (or anaphor) has been zeroed, common in pro-drop languages. The first corpus of this kind was described in (Pereira, 2009). An English-Portuguese comparable corpus was used in a study to resolve coreference in dialogues in (Pereira, 2009).

To our knowledge, there are no further coreference-annotated corpora containing the four languages that we include into ParCorFull2.0: English, German, French and Portuguese.

3. Annotation Categories

The annotation of the French and Portuguese texts is based on the annotation guidelines by Lapshinova-Koltunski and Hardmeier (2017). They address the segmentation of nominal elements, the annotation of different antecedent and anaphora types and examples of various problematic cases. The specific properties of French and Portuguese required adaptation of the annotation guidelines. Both of these languages contain clitic pronouns. This implies different mention selection strategies in MMAX2 (see below). The Portuguese texts were annotated using the guidelines for French. However, we also introduced a new category here which was not available in the other languages at hand – null or zero anaphors for pro-drops.

As a reminder, we include an overview of the main annotation principles in ParCorFull.

Segmentation The annotated elements (markables) include: pronouns, nouns, noun phrases or elliptical constructions that are parts of a coreference pair (antecedent-anaphora), as well as verb phrases or clauses being antecedents of event anaphora.

Types of antecedents We include both entities and events as antecedents. Entities can either be represented by a pronoun or a noun phrase. Events can be represented by a VP, a clause or a set of clauses, see example (3) in (Lapshinova-Koltunski et al., 2018). Antecedents can be split, and if there is no explicit antecedent, the position of the antecedent is left open. The latter occurs if a referring expression is anaphoric, but no specific antecedent can be found in the text.

Types of anaphora Most referring expressions (anaphors) are constituted by pronouns and nominal phrases are annotated as referring expressions (anaphors). However, in some cases, referering expressions are parts of verbs or verbal phrases, e.g. in case of elliptical constructions or clitics in Romance languages (see below).

Coreferring **pronouns** include demonstrative, personal, relative, reflexive pronouns and the category none (for pro-drops as stated below). In French, we annotate the indirect pronouns *y* and *en*, see example (2-a), except when they occur as fixed elements of constructions like *il y a* 'there is' or *s'en aller* 'to leave' (2-b).

(2) a. *Jean est allé à [Paris]. Il [y] a trouvé le bonheur.* ("Jean has gone to Paris. He's found happiness there.")
b. *Jean est allé à [Paris]. Il y a beaucoup de monde à [Paris].* ("Jean has gone to Paris. There are a lot of people in Paris.")

Relative pronouns in Portuguese may contain further elements (e.g. article or a preposition). In this case, the whole relative pronoun phrase is marked, as shown in example (3-a) where *de que* is marked as a referring expression. In French, the relative pronouns *ce qui*, *ce* *que* and *ce dont* similarly consist of two words annotated as a single unit (3-b). These elements are not annotated when they are used as interrogative pronouns.

(3) a. *Mas [a ideia] [de que] não devemos permitir que a ciência faça o seu trabalho porque temos medo, é de facto muito sufocante.* ("But [the idea] [that] we should not allow science to do its job because we're afraid, is really very deadening").

b. *Le docteur m'a dit que j'étais guéri, [ce qui] m'a surpris.* ("The doctor told me I was cured, [which] surprised me.")

c. *... alors que tu es en train d'étudier tout ce qui peut mal se passer* ("... while you're studying *all that* can go wrong.")

In addition to pronouns, we also mark up deictic adverbs pointing to locations (*there, here*) and moments in time (*then, now*) if they have an identifiable antecedent in the text.

Personal pronouns functioning as direct objects and reflexives in Portuguese are frequently used as clitic pronouns, i.e. they are joined to the head word – the verb. For example, see the verbs *auto-identificarem-se* and *associarem-se* in example (1) above. Our annotation strategy is to mark the whole verb which is concatenated with the pronominal element, as MMAX2 does not allow their separate marking. Another strategy would be a different tokenisation – separation of clitics in the pre-annotation step. However, we consider marking clitics together with verbs for a better and more convenient option. Although the verb merged with the clitic is marked, such cases are annotated as pronouns and not verbs. For instance, *convidá-los* with the clitic pronoun -los in (4) is marked as a pronouns, which has an anaphoric functions, referes to an entitiy (*universidades de toda a África subsaariana*), which can be characterised as a simplae antecedent. Besides that, this is a personal pronoun in plural with non-subject function.

(4) *Trabalhamos com [universidades de toda a África subsaariana] e estamos a [convidá-los] a adquirir competências em inovação social.* ("Working with [universities all over sub-Saharan Africa], And we are inviting [them] to learn social innovation skills. ").

Both Portuguese and French belong to clitic-doubling languages, i.e. pronominal reduplication may occur here, see example (5), with *se* and *si*. Unlike Spanish, where clitic pronouns may double full nominal phrases, both in Portuguese and French, they can double pronouns only (Magro, 2019). However, such cases are not so frequent in our corpus.

(5) *a possibilidade de [um indivíduo] [se] ver a [si] próprio como capaz.* ("the possibility of [an individual] (themselves) to see [themselves] as capable").

As we cannot technically mark zeros, i.e. in case of null or zero anaphors or pro-drops, we also mark the gead verbs, see the verb *quiserem* in example (1) above. This category does not exist the other languages of our corpus and was not originally forseen in our annotation scheme. However, the results from our preliminary corpus analyses showed a certain degree of loss in the data, if this category was not considered: the Portuguese texts contained much fewer pronominal anaphors as their English and German counterparts. Instead of adding a new category into the scheme, we used the option 'none' already encoded in the MMAX2 scheme. Similarly as with the case of clitics, we mark the head verb but annotate it as a pronoun. For instance, *quiserem* in example (1) above, is annotated as a pronoun with th following features: anaphoric, plural, subject and none, which refers to an entity being a simple antecedent.

Coreferring **noun phrases** include proper names (*Simone Biles* in Figure 1 below), nominal postmodifiers as *Ivanov* in the nominal phrase *O adjunto de Ivanov* in (6-a), full noun phrases as *uma empresa com o Stan Winston* in example (6-b). Generic nouns like *A acção colectiva* in (6-c) can co-refer with definite full NPs or pronoun incuding zero anaphors (*coloca* and *causa*), but not with other generic nouns.

(6) a. *O adjunto de [Ivanov] desde 2012, Anton Vaino, foi nomeado como [seu] sucessor.* ("Mr [Ivanov]'s deputy since 2012, Anton Vaino, has been appointed as [his] successor.")

b. *Então , criei [uma empresa com o Stan Winston], (...) E o conceito [da empresa] era...* ("So , I started [a company with Stan Winston]... And the concept of [the company] was...")

c. *[A acção colectiva] pouco ou nada consegue, mas [coloca] pressão sobre equipas e serviços já sobrecarregados e [causa] preocupação...* ("[Industrial action] achieves little or nothing, but [it] places pressure on already stretched teams and services and [it] causes worry")

Linguistic chains may also include *substitution* and *ellipsis* in addition to referring expressions. These trigger a type reference relation (as opposed to a relation of identity) between referents belonging to the same class (Kunz and Steiner, 2013; De Beaugrande and Dressler, 1981). In substitution patterns, the referring expression is replaced with another element (see example of verbal substitution in (8) below). In ellipsis, it is completely left out, and the reference is implicit, as in example (7).

(7) *Eu pedi a alguém que contasse o número de [livros com felicidade no título , publicados nos últimos cinco anos], e eles desistiram depois*

de cerca de 40, e havia muitos mais []. ("I had somebody count the number of books with "happiness" in the title published in the last five years and they gave up after about 40, and there were many more [].").

We include substitution and ellipsis into our framework, since they often occur in similar contexts as coreference if considered cross-lingually. We subdivided them into their structural types, according to the omitted/substituted element: nominal, verbal and clausal. In example (8), we illustrate a case of a clausal ellipsis[2] with *Não* ("No") and a verbal substitution with *Fazem-no* ("do so").

(8) *...Miguel, [elas voam 240 km até à propriedade e depois voam 240 km de volta à noite]? [Fazem-no] pelas crias?... [Não], respondeu. [Fazem-no] porque a comida é melhor.* (...Miguel, do [they fly 150 miles to the farm, and then do they fly 150 miles back at night]? Do they [do so] for the children?... [No]. They [do so] because the food's better.")

Another category that is considered here but is excluded from most analyses[3] is that of comparative reference, which does not trigger co-reference in the strict sense. Together with other cases (substitution and ellipsis) it instead involves type reference, co-classification or "sloppy identity" (Kunz and Steiner, 2012). The linguistic means signaling comparative reference include such words as *same, equal, identical* or particular adjectives in the comparative form. We distinguish between general and particular comparison, the first referring to a general relation of comparison between two entities (9-a) and the latter referring to particular comparative features of two entities (9-b).

(9) a. *Centenas de milhares de mortes desnecessárias num [país] que tem sido atormentado mais do que [qualquer outro], por esta doença.* ("Hundreds of thousands of needless deaths in [a country] that has been plagued worse than [any other] by this disease.")

b. *That car over there is very [fast]. But well, my uncle drives an even [faster] one.*

4. Data Selection

We extend ParCorFull with French and Portuguese translations of a subset of the texts already included in the data. For French, we were able to add the complete set of 20 TED talks present in the English and German subcorpora. Due to project priorities, no news data was added for French[4].

For the Portuguese subcorpus, we extracted 11 TED talks from the Portuguese part of the 2017 IWSLT evaluation campaign. They are translations of the same 11 English TED talks that were included into ParCorFull (derived originally from the ParCor corpus (Guillou et al., 2014)) and correspondingly, they are parallel with the English sources and their translations into German. For nine of the talks included in the English, German and French subset, no Portuguese translation was available. The WMT news test sets of the news translation shared task at the Conference on Machine Translation (Bojar et al., 2017, WMT2017) do not contain Portuguese translations either. However, we included a part of the news texts translated by a translation office for our project.

As a result, the ParCorFull2.0 corpus contains a common subset of 11 TED talks fully annotated and parallel across all four languages, English, German, French and Portuguese. Nine more TED talks are included in English, German and French only. The news portion contains 11 articles available in English, German and Portuguese and 8 additional articles in English and German only.

Table 1 provides an overview of the total number of texts and tokens for all languages contained in the current version of the corpus.

5. Annotation Process

All the annotations were perfomed with the help of the annotation tool MMAX2 (Müller and Strube, 2006). The annotation scheme created for this task allows human annotators to define each markable as a certain mention type (pronoun, NP, VP or clause). Then, the mentions can be defined further in terms of their cohesive function (antecedent, anaphoric, cataphoric, comparative, etc.). Antecedents can either be annotated as simple or split, and as entity or event. For anaphoric expressions the scheme includes singular/plural agreement with the antecedent and subject/non-subject position of the expression. The annotation scheme also covers pronoun type (personal, possessive, demonstrative, reflexive, relative and none for zero anaphors) and modifier types of NPs (possessive, demonstrative, definite article, or none for proper names). An example of the MMAX2 interface with a visualisation of a coreference chain in Portuguese is illustrated in Figure 1. Annotations for all languages were performed by highly experienced well-trained annotators with linguistic background in order to ensure maximum accuracy.

6. Annotation Results

Table 2 presents an overview of the annotated structures (in absolute numbers).

The current corpus version contains about 28,000 annotated mentions at the moment (counted for all languages). We group the annotated mentions according to their morpho-syntactic type in Table 2: pronouns

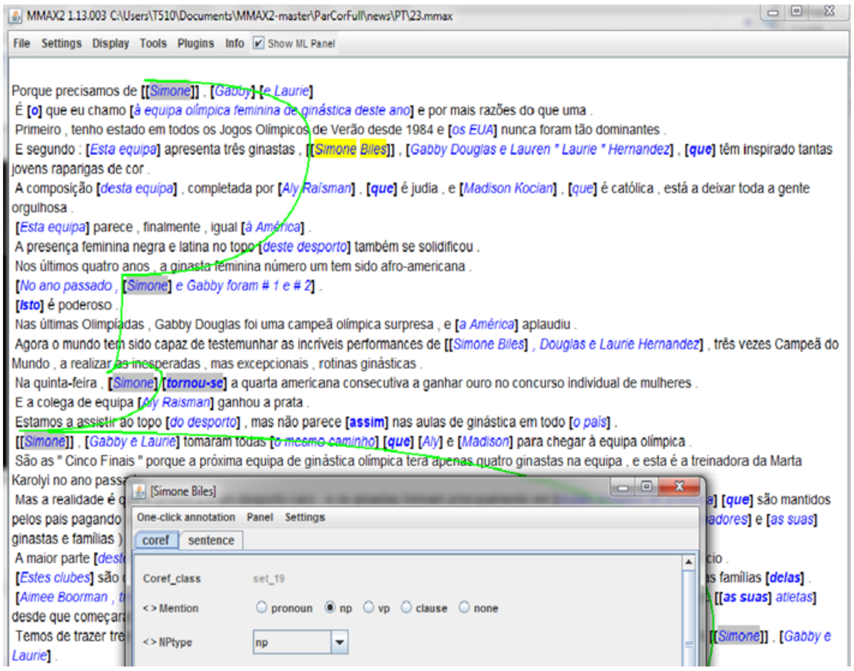
[2] Also called yes-no-ellipsis by Menzel (2017)

[3] The only computational study known to us is (Khullar et al., 2020) on what the authors call one-anaphora.

[4] The WMT news data does not contain French.

	TED Talks			News			Total		
language	txt	snt	token	txt	snt	token	txt	snt	token
English	20	3,277	70,736	19	464	10,798	39	3,741	81,534
German	20	2,829	66,783	19	281	10,602	39	3,110	77,385
French	20	1,959	76,229	–	–	–	20	1,959	76,229
Portuguese	9	1,488	27,898	11	309	6,522	20	1,797	34,420
Total	69	9,553	241,646	49	1,054	27,922	118	10,607	269,568

Table 1: Statistics on the corpus data: number of texts (txt), sentences (snt) and number of tokens (token).


MMAX2 1.13.003 C:\Users\T510\Documents\MMAX2-master\ParCorFull\news\PT\23.mmax
File
Settings
Display
Tools
Plugins
Info
☑ Show ML Panel

Porque precisamos de [[Simone]] . [Gabby] [e Laurie]
É [o] que eu chamo [à equipa olímpica feminina de ginástica deste ano] e por mais razões do que uma .
Primeiro , tenho estado em todos os Jogos Olímpicos de Verão desde 1984 e [os EUA] nunca foram tão dominantes .
E segundo : [Esta equipa] apresenta três ginastas , [[Simone Biles]] , [Gabby Douglas e Lauren " Laurie " Hernandez] , [que] têm inspirado tantas jovens raparigas de cor .
A composição [desta equipa] , completada por [Aly Raisman] , [que] é judia , e [Madison Kocian] , [que] é católica , está a deixar toda a gente orgulhosa .
[Esta equipa] parece , finalmente , igual [à América] .
A presença feminina negra e latina no topo [deste desporto] também se solidificou .
Nos últimos quatro anos , a ginasta feminina número um tem sido afro-americana .
[No ano passado , [Simone] e Gabby foram # 1 e # 2] .
[Isto] é poderoso .
Nas últimas Olimpíadas , Gabby Douglas foi uma campeã olímpica surpresa , e [a América] aplaudiu .
Agora o mundo tem sido capaz de testemunhar as incríveis performances de [[Simone Biles] , Douglas e Laurie Hernandez] , três vezes Campeã do Mundo , a realizar as inesperadas , mas excepcionais , rotinas ginásticas .
Na quinta-feira , [Simone] [tornou-se] a quarta americana consecutiva a ganhar ouro no concurso individual de mulheres .
E a colega de equipa [Aly Raisman] ganhou a prata .
Estamos a assistir ao topo [do desporto] , mas não parece [assim] nas aulas de ginástica em todo [o país] .
[[Simone]] , [Gabby e Laurie] tomaram todas [o mesmo caminho] [que] [Aly] e [Madison] para chegar à equipa olímpica .
São as " Cinco Finais " porque a próxima equipa de ginástica olímpica terá apenas quatro ginastas na equipa , e esta é a treinadora da Marta Karolyi no ano pass
Mas a realidade é q … a] [que] são mantidos pelos pais pagando … adores] e [as suas] ginastas e famílias) …
A maior parte [dest … cio .
[Estes clubes] são … s famílias [delas] .
[Aimee Boorman , t … [[as suas] atletas] desde que começar …
Temos de trazer tre … [[Simone]] , [Gabby e Laurie] .

Coref_class set_19
<> Mention ○ pronoun ● np ○ vp ○ clause ○ none
<> NPtype np

Figure 1: A coreference chain in a Portuguese news text visualised in MMAX2.

	English			German			French	Portuguese			total
	news	TED	total	news	TED	total	TED	news	TED	total	total
pron	400	3,772	4,172	477	3,840	4,317	5,140	329	1,772	2,101	15,730
np	434	2,206	2,640	446	2,401	2,847	3,327	410	1,501	1,911	10,725
vp	6	126	132	9	126	135	182	15	104	119	568
clause	12	323	335	18	317	335	360	11	127	138	1,168
all	852	6,427	7,279	950	6,684	7,634	9,009	765	3,504	4,269	28,191

Table 2: Statistics on the annotated mentions and their subcategories: pronouns (pron), nominal phrases (np), verbal phrases (vp), clauses.

(pron), nominal phrases (np), verbal phrases (vp) and clauses (clause). This differentiation was introduced for a practical reason, as it permits classifying mentions further according to their function or the role in a coreference chain. The numbers in the table reveal that pronominal mentions are most frequent in all languages (although their number is not much higher than that of nominal mentions in Portuguese). We also see that the French data contains many more mentions than the other languages. The Portuguese data is the smallest due to a smaller number of texts and tokens (see Table 1 above).

The number of full coreference chains in the data amounts to 10,696 (see Table 3). We also calculate the average chain length. The French translations contain much more chains than their English sources and other translations. It is also interesting to see that the German and the French translations contain shorter chains than the English originals, whereas the Portuguese translations contain longer chains on average.

Computing reliable inter-annotator agreement scores for French turned out to be difficult due to the the his-

	nr. chain	chain/snt	**av. length**
English	2,319	0.62	2.94
German	2,425	0.78	2.81
French	4,744	2.42	2.87
Portuguese	1,208	0.67	3.22
total	10,696	1.00	-

Table 3: Statistics on the annotated chains: total number of chains (nr. chain), chains per sentence (chain/snt) and average chain length (av.length).

tory of the corpus creation. The annotation began in 2018 with one annotator, who dropped out after annotating a small number of texts. We subsequently identified significant quality problems in those texts and decided to restart the complete annotation with a new annotator, who also revised the initially annotated texts. Comparing the annotations of the first and the second annotator, we find a mention identification F-score of 81.9% and a CEAF*e* score of 72.7%. These scores quantify the number of changes required by the reannotation, but do not adequately reflect the annotation difficulty since the second annotator had access to the first annotator's output while completing her work. In a later attempt of creating an ad-hoc second annotation, we observed much lower agreement scores of 63.7% (mention identification) and 50.7% (CEAF*e*). A closer study of the discrepancies between the annotations revealed that the vast majority of them were due to the double annotator's lack of training, with the main annotation being correct in almost every case.

For Portuguese, we achieve an F-score of 83.3% in mention identification, and a CEAF*e* score of 78.3%. A qualitative analysis of differences also reveal some difference in the mention span, e.g. inclusion of the full verbal phrase instead of marking verbs only (as event antecedents). Also, even though the second annotator did not include the zero anaphors, s/he annotated more mentions. There were also differences in clustering specific mentions into chains.

7. Conclusion and Future Work

Cross-lingual differences in the realisation of coreference relation are of interest for contrastive linguists and researchers in translation studies. At the same time, they pose a challenge to multilingual natural language processing, such as machine translation or multiligual information extraction. A parallel corpus with coreference annotations in four languages is a valuable resource, which can find application in various areas. On the one hand, the corpus should help to study the mechanisms involved in coreference translation in order to develop a better understanding of the phenomenon as it was done for English in German by Lapshinova-Koltunski et al. (2020). It may also serve as a resource for creating and evaluating coreference-aware MT systems, see for instance (Lapshinova-Koltunski et al., 2019; Guillou et al., 2018), without having to rely on notoriously inaccurate automatic coreference resolvers. This corpus can also be used as part of training and development resource for the creation of multilingual or monolingual coreference resolution systems.

8. Acknowledgements

Christian Hardmeier and Elina Lartaud were supported by the Swedish Research Council under grant 2017-930, which also funded the annotation work of the French data. We also wish to thank Miryam de Lhoneux and Marie Dubremetz for their work on the French annotations. Pedro Augusto Ferreira was supported by FCT, Foundation for Science and Technology, Portugal, under grant SFRH/BD/146578/2019.

9. Bibliographical References

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

De Beaugrande, R.-A. and Dressler, W. U. (1981). *Einführung in die Textlinguistik*. Niemeyer, Tübingen.

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (2018). A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels, October. Association for Computational Linguistics.

Khullar, P., Bhattacharya, A., and Shrivastava, M. (2020). Finding the right one and resolving it. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 132–141, Online, November. Association for Computational Linguistics.

Kunz, K. and Lapshinova-Koltunski, E. (2015). Cross-linguistic analysis of discourse variation across registers. *Special Issue of Nordic Journal of English Studies*, 14(1):258–288.

Kunz, K. and Steiner, E. (2012). Towards a comparison of cohesive reference in English and German: System and text. In M. Taboada, et al., editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*. Equinox, London.

Kunz, K. and Steiner, E. (2013). Cohesive substitution in English and German: A contrastive and corpus-based perspectivet. In Karin Aijmer et al., editors, *Advances in Corpus-Based Contrastive Linguistics. Studies in honour of Stig Johansson*, pages 201–232. John Benjamins, Amsterdam.

Lapshinova-Koltunski, E. and Hardmeier, C., (2017). *Coreference Corpus Annotation Guidelines*, December.

Lapshinova-Koltunski, E. and Kunz, K. (2020). Exploring coreference features in heterogeneous data. In *1st Workshop on Computational Approaches to Discourse (CODI-2020)*, pages 53–64. ACL, 20 November. EMNLP workshop.

Lapshinova-Koltunski, E., España-Bonet, C., and van Genabith, J. (2019). Analysing coreference in transformer outputs. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019) at EMNLP-2019*, pages 1–12, Hong Kong, China, November 3. Association for Computational Linguistics.

Lapshinova-Koltunski, E., Krielke, M.-P., and Hardmeier, C. (2020). Coreference strategies in English-German translation. In *3rd Workshop on omputational Models of Reference, Anaphora and Coreference (CRAC-2020)*, pages 139–153. COLING, 12 December. COLING workshop.

Magro, C. (2019). Redobro de clítico em português europeu. *Estudos de Lingüística Galega*, 11:29–75.

Menzel, K. (2017). *Understanding English-German contrasts: a corpus-based comparative analysis of ellipses as cohesive devices*. Ph.D. thesis, Universität des Saarlandes, Saarbrücken.

Naumann, K. and Möller, V. (2007). Manual for the annotation of in-document referential relations. Technical report, University of Tübingen, May.

Novák, M. and Nedoluzhko, A. (2015). Correspondences between Czech and English coreferential expressions. *Discours*, 16.

10. Language Resource References

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain, April. Association for Computational Linguistics.

Aktaş, B. and Kohnert, A. (2020). TwiConv: A coreference-annotated corpus of Twitter conversations. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 47–54, Barcelona, Spain (online), December. Association for Computational Linguistics.

Bourgonje, P. and Stede, M. (2020). The Potsdam Commentary Corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

Collovini, S., Carbonel, T., Fuchs, J., Coelho, J., Rino, L., and Vieira, R. (2007). Summ-It: Um corpus anotado com informações discursivas visando a sumarização automática. In *Proceedings of V Workshop Em Tecnologia Da Informação e Da Linguagem Humana*, pages 1605—1614, Rio de Janeiro.

da Fonseca, E. B., Sesti, V., Collovini, S., Vieira, R., Leal, A. L., and Quaresma, P. (2017). Collective elaboration of a coreference annotated corpus for Portuguese texts. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, Murcia, Spain. co-located with 33th Conference of the Spanish Society for Natural Language Process.

Garcia, M. and Gamallo, P. (2014). Multilingual corpora with coreferential annotation of person entities. pages 3229—3233.

Gardent, C. and Manuelian, H. (2005). Création d'un corpus annoté de traitement des descriptions définies. *Traitement Automatique des Langues (TAL)*, 46:115–140.

Grishina, Y. and Stede, M. (2015). Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, page 14, Beijing, China.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland.

Kunz, K., Lapshinova-Koltunski, E., Menzel, K., Steiner, E., and Martínez, J. M. M. (2021). *GECCo – German-English Contrasts in Cohesion*, volume 355 of *Trends in Linguistics. Studies and Monographs [TiLSM]*. Mouton de Gruyter.

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de réf'erence (democrat). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, 92:11–15.

Lapshinova-Koltunski, E., Hardmeier, C., and Krielke, P. (2018). ParCorFull: a Parallel Corpus Annotated with Full Coreference. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of 11th Language Resources and Evaluation Conference*, pages 423–428, Miyazaki, Japan, may. European Language Resources Association (ELRA).

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, et al., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Nedoluzhko, A., Novák, M., Cinková, S., Mikulová, M., and Mírovský, J. (2016). Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 169–176, Portorož, Slovenia. European Language Resources Association.

Pereira, S. (2009). ZAC.PB: An annotated corpus for zero anaphora resolution in Portuguese. In *Proceedings of the Student Research Workshop*, pages 53–59, Borovets, Bulgaria, September. Association for Computational Linguistics.

Todirascu, A. (without year). Le corpus evalrefgen. `https://groupes.renater.fr/wiki/corpus-ecrits/_media/public/corpusevalrefgen.pdf`. Accessed: 2022-01-04.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., and Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC 2000)*, Lancaster, UK.

Uryupina, O., Artstein, R., Bristot, A., Cavicchio, F., Delogu, F., Rodriguez, K. J., and Poesio, M. (2020). Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

Vieira, R., Mendes, A., Quaresma, P., Fonseca, E., Collovini, S., and Antunes, S. (2018). Corref-PT: A semi-automatic annotated portuguese coreference corpus. *Computación y Sistemas*, 22:1259—1267.

Weischedel, R. M., Hovy, E. H., Marcus, M. P., and Palmer, M. (2017). OntoNotes : A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63, New York. Springer-Verlag.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51:581–612.