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

Setting Goals and Motivation

Contrastive Example

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.*

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.*

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.*

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.*

Goals

- ParCorFull: **full coreference** ⇒ pronouns, full nominal and verbal phrases and clauses.
- ParCorFull: EN→DE, ParCorFull2.0: **EN→DE, EN→FR and EN→PT**.

Motivation

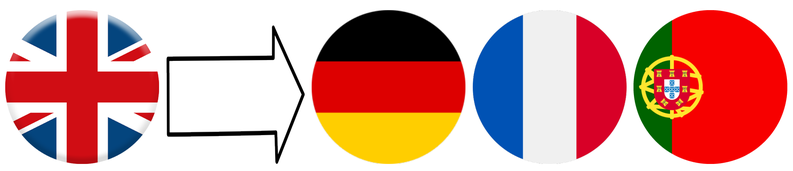
- Coreference across languages: **a problem** for MT and other multilingual NLP technologies.
- Choice between referring expressions is governed by **language-specific constraints**.
- FR and PT pose **new annotation challenges**.

Existing Resources

- EN: ARRAU, GUM, OntoNotes, PCEDT, Czech-PCEDT, TwiConv, Parallel Meaning Bank.
- DE: TüBa/DZ, PotsdamCC, ParCorFull, ParCor, GECCo.
- FR: DeDe, ANCOR, Democrat, EvalRefGen, ELRA-W0032.
- PT: HAREM, Summ-it, Corref-PT, ZAC

Data and Annotation Principlese

Texts (news from WMT and TED talks) in multiple languages:

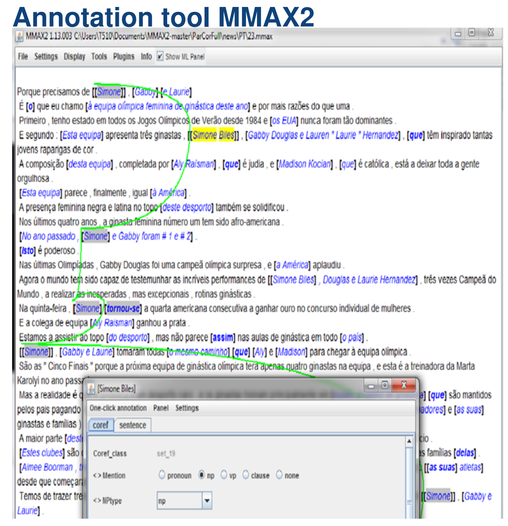


	TED Talks			News			Total		
language	txt	snt	token	txt	snt	token	txt	snt	token
English	20	3,277	70,736	19	464	10,798	39	3,741	81,534
German	20	2,829	66,783	19	281	10,602	39	3,110	77,385
French	20	1,959	76,229	–	–	–	20	1,959	76,229
Portuguese	9	1,488	27,898	11	309	6,522	20	1,797	34,420
Total	69	9,553	241,646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

Annotation categories

ANNOTATION	MARKABLES:			
Segmentation:	PRONOUN	NP	VP	CLAUSE
Type:	Pers. Rel. Refl. Dem. Pos. (incl. **locations** & **time** (then, now) & **pronominal adverbs**	Bare Poss. Def. Art. Dem. Ellipsis Apposition	Subst. Ellipsis	Subst. Ellipsis
Function as:	ANAPHORA* / ANTECEDENTS		ANTECEDENTS	
Types of antecedents:	ENTITIES (Simple, split, no explicit)		EVENTS	

Annotation tool MMAX2



Examples

Indirect pronouns:

(1) *Jean est allé à Paris. Il y a trouvé le bonheur.* ("Jean has gone to Paris. He's found happiness there.")

Relative pronouns with further elements:

(2) 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.")

Clitics:

(3) *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. ").

Double clitics:

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

Nominal phrases:

(5) 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...")

Ellipsis:

(6) *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 [].").

Clausal ellipsis and verbal substitution:

(7) *. . . 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.")

Comparative reference:

(8) 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.*

Annotation results

	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**

	nr. chain	chain/snt	av. length
EN	2,319	0.62	2.94
DE	2,425	0.78	2.81
FR	4,744	2.42	2.87
PT	1,208	0.67	3.22
total	**10,696**	1.00	-

	mention	CEAFe
FR	81.9%	72.7%
PT	83.3%	78.3%

Summary

FR contains much more chains and mentions.
DE and FR contain shorter chains, whereas PT contains longer chains on average.
Pronominal mentions are most frequent in all languages.

Corpus availability

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

Acknowledgements

Christian Hardmeier and Elina Lartaud were supported by the Swedish Research Council under grant 2017-930. We thank Miryam de Lhoneux and Marie Dubremetz for their work on the French annotations. Pedro Augusto Ferreira was supported by FCT under grant SFRH/BD/146578/2019.