MultiSub: A multiple parallel subtitle corpus

Fahime Same [1], **Laura Becker** [2], **Alessia Cassarà** [1]
f.same@uni-koeln.de, gombos.becker@fau.de, alessia.cassara@uni-koeln.de
University of Cologne, University of Erlangen-Nürnberg

In a nutshell

Currently available multiple parallel corpora

- there are a number of open-access parallel corpora targeting the linguistic community: Tiedemann (2012), Lison and Tiedemann (2016), Levshina (2016), and Cysouw and Wälchli (2007)
- many either offer only language pairs and/or do not feature additional linguistic or extra-linguistic annotation

MultiSub

- multiple parallel texts (subtitles of TV series)
  subtitles from `http://www.addic7ed.com/`
- linguistic & extra-linguistic annotation
- naturalistic data

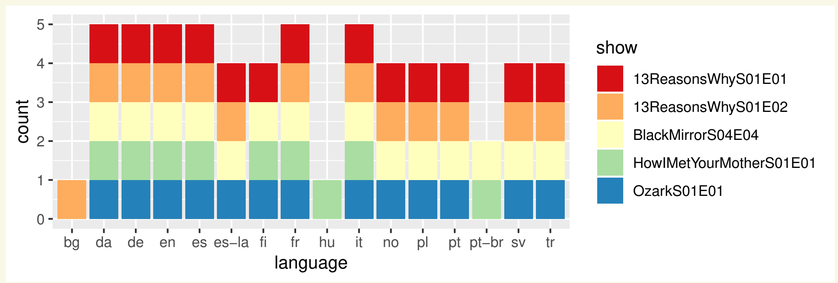

Figure 1: Shows and languages available so far

Processing cycle

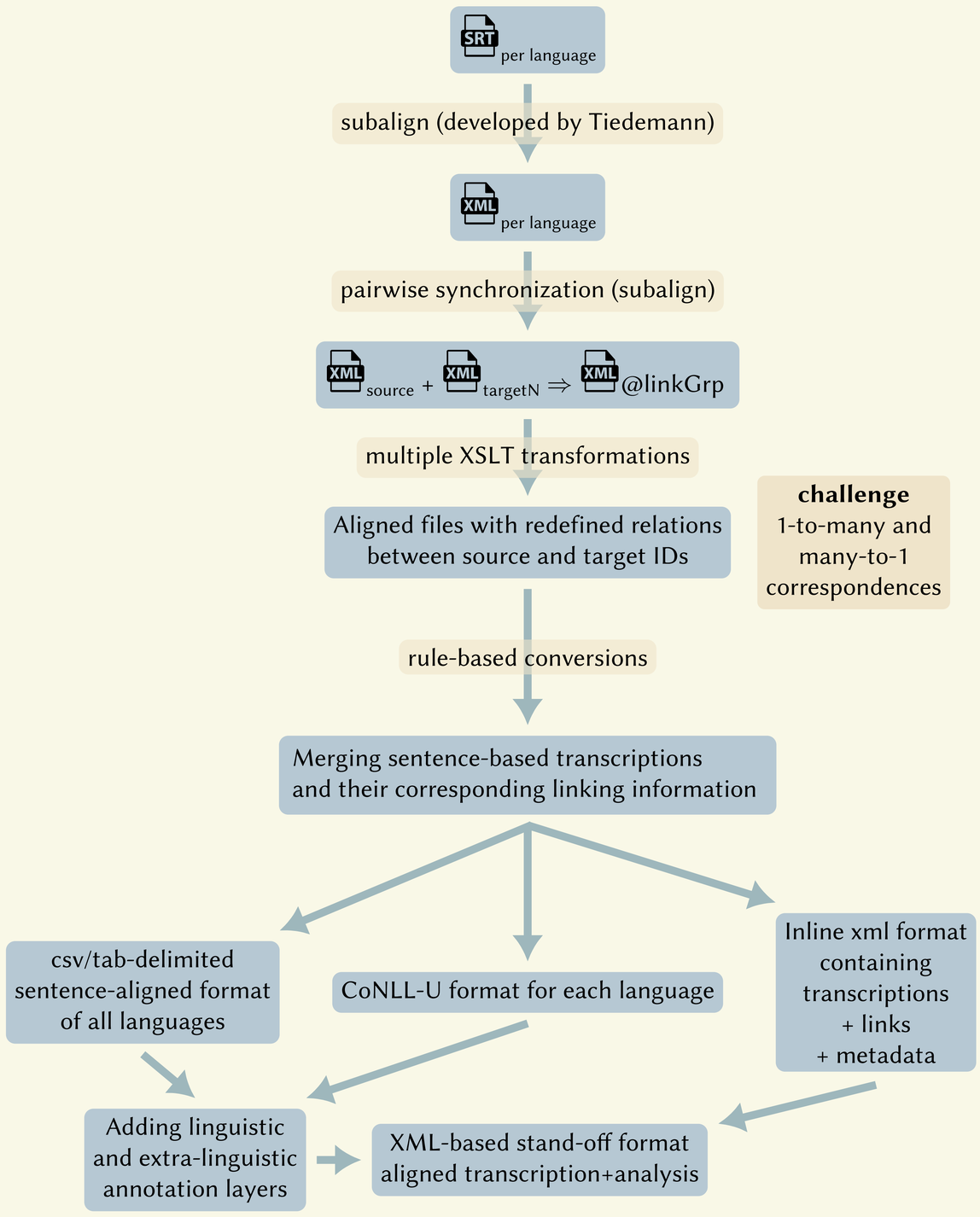

Figure 2: Workflow

Extra-linguistic annotation (manual)

Speaker information

- important for the study of turn taking, conversation structuring
- important for the comparison of speaker-anaphora vs. hearer-anaphora

Scene transitions

- a scene is defined as a setting without any noticeable break in the ongoing stream of time (in line with Häusler and Hanke 2016)
- scene-cuts are based on temporal progression
- important for the study of discourse, referential variation, etc.

check out the multisub project website: `https://osf.io/8wp4q/`

Linguistic annotation (automatic)

Tool: UDPipe (Straka and Straková 2017)

Analysis: lemmatization, POS tagging, dependency parsing (s. Table 2)

Advantage: allows for direct, crosslinguistic structural comparison

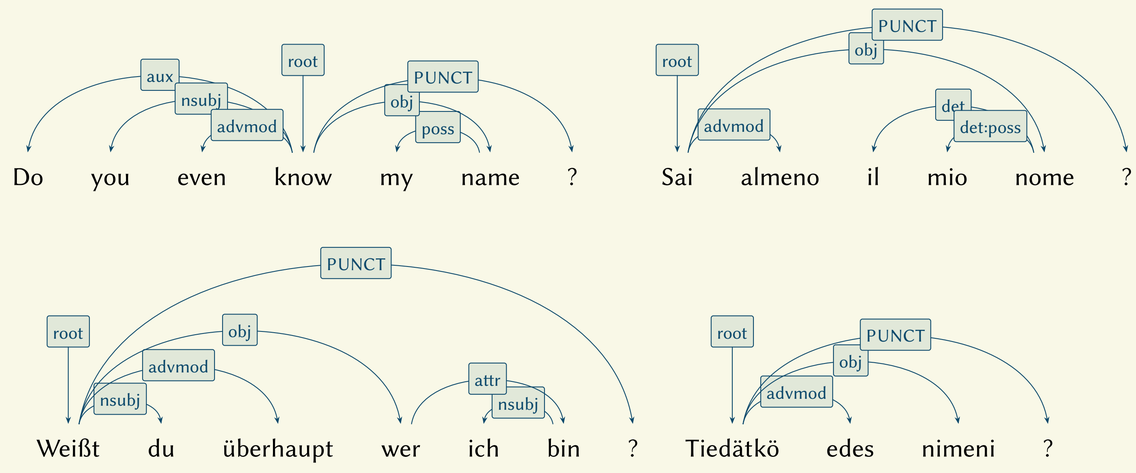

Figure 3: Dependency structure for the same sentence in 4 languages

Output formats

1. Multi-language parallel utterance-based segmentation

scene	speaker	id-en	en	id-de	de	id-fr	fr	id-sp	sp
4	Marty	278	Is she all right?	227	- Geht es ihr gut?	223	- Qu'est- ce qu'elle a?	212	- ¿Le pasa algo?
4	Wendy	279	- She has psoriasis.	228	- Sie hat Schuppenflechte.	224	- Du psoriasis.	213	- Tiene psoriasis.
4	Marty	280	- Oh.	228	- Sie hat Schuppenflechte.	225	- C'est terrible.	213	- Tiene psoriasis.
4	Marty	281	Jesus Christ.	229	- Meine Güte.	225	- C'est terrible.	214	- Joder.
4	Charlotte	282	It's a disease, Dad.	230	Das ist eine Krankheit, Dad .	226	C'est une maladie, papa.	215	Es una enfermedad, papá.
4	Charlotte	283	Like cancer, okay?	231	So wie Krebs, ok?	227	Comme le cancer.	216	Como el cáncer, ¿ vale?
4	Marty	284	No.	232	Nein.	228	Non.	217	No.
4	Marty	285	Is it?	233	Echt?	229	Tu crois?	218	¿ O sí?
4	Marty	286	It's itchy skin, honey.	234	Das ist juckende Haut.	230	Elle a des démangeaisons.	219	Es picor de piel.
4	Charlotte	287	There's no cure.	235	Es gibt kein Heilmittel.	231	C'est incurable.	220	No tiene cura.
4	Marty	288	Right, let's save our money.	236	Ok, sparen wir unser Geld.	232	Alors, gardons notre argent.	221	Vale, ahorremos el dinero.

Table 1: Tab-delimited format with extra-linguistic information

2. CoNLL-U format

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	MISC
1	Acho	achar	VERB	VERB	_	0	root	22-pt-17.1
2	que	_	CCONJ	CONJ	_	7	mark	22-pt-17.2
3	a	_	DET	DET	_	4	det	22-pt-17.3
4	maioria	maioria	NOUN	NOUN	_	7	nsubj	22-pt-17.4
5	das	_	X	ADPPRON	_	4	appos	22-pt-17.5
6	pessoas	pessoa	NOUN	NOUN	_	5	flat	22-pt-17.6
7	tem	ter	VERB	VERB	_	1	ccomp	22-pt-17.7
8	uma	_	DET	DET	_	9	det	22-pt-17.8
9	visão	visão	NOUN	NOUN	_	7	obj	22-pt-17.9
10	errada	errado	ADJ	ADJ	_	9	amod	22-pt-17.10
11	do	_	X	ADPPRON	_	12	case	22-pt-17.11
12	dinheiro	dinheiro	NOUN	NOUN	_	9	nmod	22-pt-17.12
13	.	_	PUNCT	.	_	1	punct	22-pt-17.13

Table 2: CoNLL-U format for a sentence in Portuguese

3. xml format

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
   <linkGrp>
      <link xml:id="SL1" target-en="#293" target-da="#263" target-de="
       #270"/>
   </linkGrp>
   <body xml:lang="da">
      <u xml:id="263">
         <seg>
            <w xml:id="263.1">Nej</w>
            <w xml:id="263.2">,</w>
            <w xml:id="263.3">ikke</w>
            <w xml:id="263.4">rigtigt</w>
            <pc xml:id="263.5">.</pc>
         </seg>
         <seg xml:id="s-da-263" xml:type="s"> Nej , ikke rigtigt .</seg>
      </u>
   </body>
   <!-- separate body for each language -->
</document>
```

Outlook & Applications

Next steps xml-based stand-off format, TEI-conformant, as well as reference annotation

Applications

- comparative discourse and syntactic analyses
- tab-delimited format: exploratory & qualitative studies, string queries
- CoNLL-U & xml format: comparative, quantitative studies including linguistic and extra-linguistic information

References

Cysouw, M. and B. Wälchli (2007). "Parallel Texts: Using Translational Equivalents in Linguistic Typology". In: *Sprachtypologie und Universalienforschung (STUF)* 60.2, pp. 95–99. Häusler, C. O. and M. Hanke (2016). "An annotation of cuts, depicted locations, and temporal progression in the motion picture" Forrest Gump"". In: *F1000Research* 5. Levshina, N. (2016). "Verbs of Letting in Germanic and Romance: A Quantitative Investigation Based on a Parallel Corpus of Film Subtitles". In: *Languages in Contrast* 16.1, pp. 84–117. Lison, P. and J. Tiedemann (2016). "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles". In: Straka, M. and J. Straková (2017). "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Vancouver, Canada: Association for Computational Linguistics, pp. 88–99. Tiedemann, J. (2012). "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey: European Language Resources Association (ELRA).