

AIMS AND MOTIVATION

Manually annotated parallel data are prone to inconsistencies of various kinds:



→ Inconsistencies are interesting indicators of:

- Annotation errors
- linguistic contrasts
- translation problems

Develop an automatical method for detecting annotation inconsistencies in parallel data for:

- Detection and correction of errors
- Detection of cohesively ambiguous situations
- Detection of language contrastive patterns in cohesion

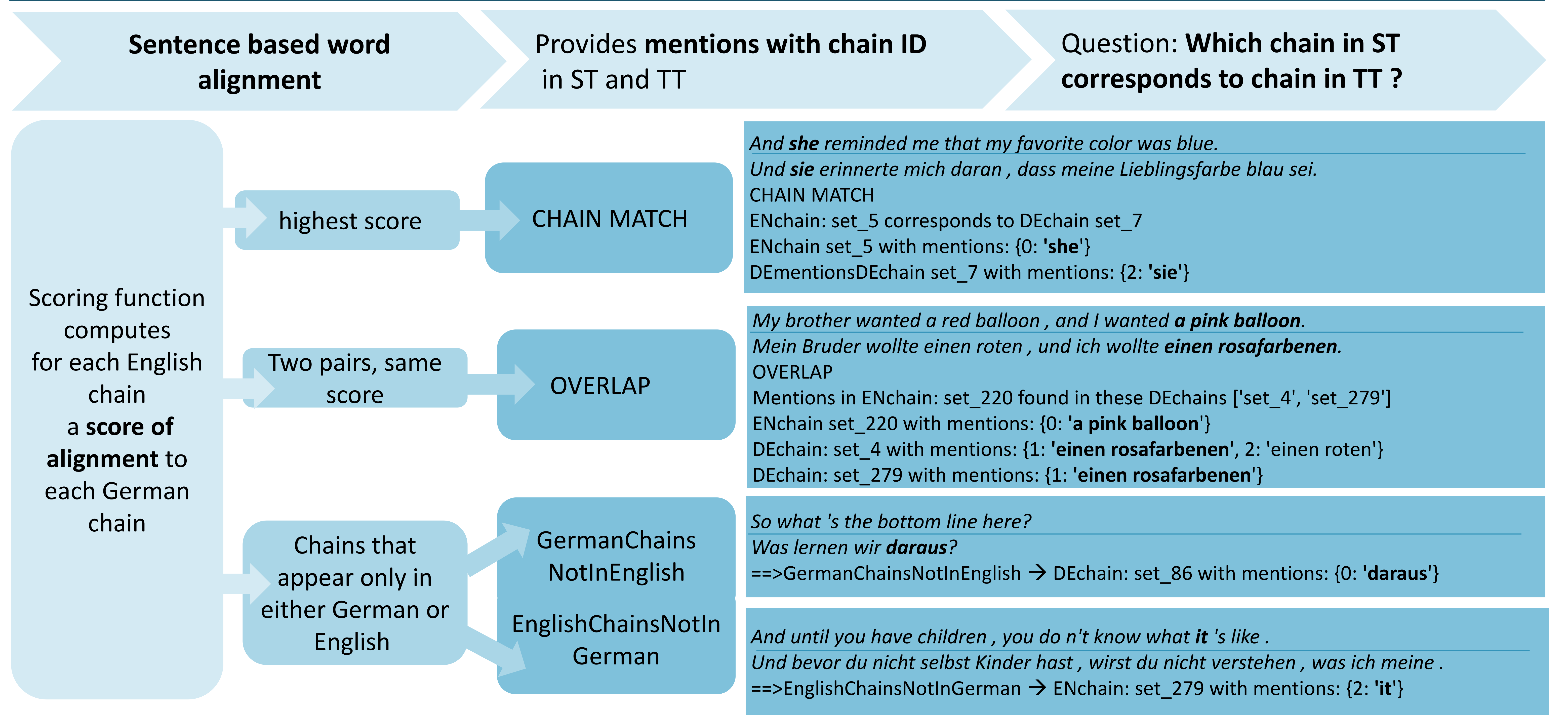
DATA

ParCor Full (Lapshinova-Koltunski, Hardmeier, Krielke 2018): Parallel corpus Annotated for FULL coreference chains Spoken 87%, written 13%. <a href="http://hdl.handle.net/11372/LRT-2614">http://hdl.handle.net/11372/LRT-2614</a>					English	German	total
	Total no of tokens	82,379	78,350	pronoun	4,650	4,269	8,919
	Total no of chains	2,319	2,425	np	2,485	2,611	5,096
	Average chain length	2.94	2.81	vp	133	132	265
				clause	335	312	647
				mentions	7,603	7,324	14,927

TYPES OF INCONSISTENCIES	...AND EXAMPLES
Linguistic	<div><div><p>"Selbst aus ganz unreligiöser Sicht <del>ist</del> <b>[die Homosexualität]</b> eine sexuelle Verfehlung dar. <b>[Sie]</b> ist ein armseliger, minderwertiger Ersatz für die Realität – ein klägliches Verstecken, dem Leben zu entfliehen. Als solches verdient <b>[sie]</b> kein Mitleid, keine Sonderbehandlung als Märtyrium einer Minderheit, und sollte nur als das betrachtet werden, was <b>[sie]</b> wirklich ist: eine schädliche Krankheit. " <b>[Das]</b> schrieb Time Magazine 1966, als ich drei Jahre alt war.</p></div><div><p>"Even in purely non-religious terms, <b>[homosexuality]</b> represents a misuse of the sexual faculty. <b>[It]</b> is a pathetic little second-rate substitute for reality – a pitiable flight from life. As such, <b>[it]</b> deserves no compassion, <b>[it]</b> deserves no treatment as minority martyrdom, and <b>[it]</b> deserves not to be deemed anything but a pernicious sickness. " <b>[That]</b>'s from Time magazine in 1966, when I was three years old.</p></div></div>
language-specific constraints	
translation-driven	
transformation patterns	
Cognitive	<div><div><p>dass <b>[es]</b> funktioniere. Und weil man <b>[diese 99 %]</b> verbrennt, ist das Kostenprofil sehr viel besser. Tatsächlich verbrennt man den Müll, und man kann sogar <b>[den Abfall von heutigen Reaktoren als Antriebsstoff]</b> benutzen. Anstatt <b>[sich] [darüber]</b> den Kopf zu zerbrechen, verbrennen Sie <b>[es]</b> einfach. Eine tolle Sache. Das Uran wird graduell verbraucht, ein bisschen wie eine Kerze. <b>[Sie]</b> sehen, dass <b>[es]</b> eine Art Säule ist, oft als "Wandernde Welle Reaktor" bezeichnet. <b>[Das]</b> löst wirklich das Treibstoffproblem. Hier ist ein Bild eines Ortes in Kentucky. <b>[Das]</b></p></div><div><p>that now you can simulate and see that, yes, with the right material's approach, <b>[this]</b> looks like <b>[it]</b> would work. And, because you're burning <b>[that 99 percent]</b>, you have greatly improved cost profile. You actually burn up the waste, and you can actually use <b>[all the leftover waste from today's reactors]</b>. So, instead of worrying about <b>[them]</b>, you just take <b>[that, it]</b>'s a great thing – <b>[it]</b> breathes this uranium as <b>[it]</b> goes along. So <b>[it]</b>'s kind of like a candle. You can see <b>[it]</b>'s a log there, often referred to as a traveling wave reactor. In terms of fuel, <b>[this]</b> really solves the problem</p></div></div>
Misinterpretation of cohesive relations	
Plain errors	
By translator	<div><p>verbrennt man den Müll, und man kann sogar <b>[den Abfall von heutigen Reaktoren als Antriebsstoff]</b> benutzen. Anstatt <b>[sich] [darüber]</b> den Kopf zu zerbrechen, verbrennen Sie <b>[es]</b> einfach. Eine tolle Sache. <b>[Das Uran]</b> wird graduell verbraucht, ein bisschen wie eine Kerze. <b>[Sie]</b> sehen, dass <b>[es]</b> eine Art Säule ist, oft als "Wandernde Welle Reaktor" bezeichnet. <b>[Das]</b> löst wirklich das Treibstoffproblem. Hier ist ein Bild eines Ortes in Kentucky. <b>[Das]</b></p></div> <div><p>because you're burning <b>[that 99 percent]</b>, you have greatly improved cost profile. You actually burn up the waste, and you can actually use as fuel <b>[all the leftover waste from today's reactors]</b>. So, instead of worrying about <b>[them]</b>, you</p></div>
By annotator	
automatic inconsistency checks for:	
(1) marked mentions outside of chains, (2) antecedents of chains not marked as first elements of chains, (3) some other error types.	
But not for potentially missing elements!	

METHOD

... and RESULTS




	Source	Target
X ChainsNotIn Y	197	76
Chains	1,102	1,114
mentions	3,108	3,248
annotated words	9,009	7,345

\*Subset of ParCorFull

FUTURE WORK

- detailed analysis and systematic description of output cases
- visualization in annotation tool

DISCUSSION



parallel vs. comparable annotation ?