Better tags give better trees – or do they?

Ines Rehbein, Hagen Hirschmann, Anke Lüdeling and Marc Reznicek

University of Potsdam, Humboldt University of Berlin

TLT-10

Outline

Parsing learner data

- Goal:
  - creating a syntactically annotated corpus of learner language
- Challenge:
  - non-canonical structures, high variability
  - unknown words (spelling errors, inflection errors, ...)
- Required:
  - robust parsing models, must be able to handle learner errors
  - domain adaptation problem?
- But: how to analyse learner language?

How to analyse learner language?

- Learner language systematically deviates from native language
- POS of a word is determined by
  - its syntactical distribution
  - its morphological marking
  - its lexical stem
- Díaz-Negrillo et al. (2010): For learner language the clues often point to diverging word classes for one token

  Example: [...] television, radio are very **subjectives** [...]

  GR-1-C-EN-041-X (Díaz-Negrillo et al., 2010, pp. 10)

- Díaz-Negrillo et al.: tripartite POS analysis to adequately describe learner language

Our approach

- Instead of parsing learner language, we parse target hypotheses (TH)
- **TH:**
  - minimal correction of learner utterances
    → parse TH and map analysis back to the learner data
- Advantage:
  - we're able to use standard NLP tools
  - we know how to analyse the data

Target hypotheses

(1)

Mnn	muss	sich	mit	diesen	Theorien	umgehen	können
[man|one]	must	oneself	with	these	theories	deal	can

aber	sind	eigentlich	sie	nicht	praxisorientiert
but	are	actually	they	not	practise-oriented

You have to be able to trade in these theories but really they are not oriented towards practise

Target hypotheses

L2 (L1)		POS	TH	TH POS	DIFF
Mnn	[man|one]				
muss	must				
sich	oneself				
mit	with				
diesen	these				
Theorien	theories				
umgehen	deal				
können	can				
aber	but				
sind	are				
eigentlich	actually				
sie	they				
nicht	not				
praxisorientiert	practice-oriented				

Target hypotheses

L2 (L1)		POS	TH	TH POS	DIFF
Mnn	[man|one]		Man		
muss	must		muss		
sich	oneself				
mit	with		mit		
diesen	these		diesen		
Theorien	theories		Theorien		
umgehen	deal		umgehen		
können	can		können		
			,		
aber	but		aber		
			eigentlich		
sind	are		sind		
eigentlich	actually				
sie	they		sie		
nicht	not		nicht		
praxisorientiert	practice-oriented		praxisorientiert		

Target hypotheses

L2 (L1)		POS	TH	TH POS	DIFF
Mnn	[man|one]		Man		CHA
muss	must		muss		
sich	oneself				DEL
mit	with		mit		
diesen	these		diesen		
Theorien	theories		Theorien		
umgehen	deal		umgehen		
können	can		können		
			,		INS
aber	but		aber		
			eigentlich		MOVT
sind	are		sind		
eigentlich	actually				MOVS
sie	they		sie		
nicht	not		nicht		
praxisorientiert	practice-oriented		praxisorientiert		

Target hypotheses

L2 (L1)		POS	TH	TH POS	DIFF
Mnn	[man|one]		Man	PIS	CHA
muss	must		muss	VMFIN	
sich	oneself				DEL
mit	with		mit	APPR	
diesen	these		diesen	PDAT	
Theorien	theories		Theorien	NN	
umgehen	deal		umgehen	VVINF	
können	can		können	VMINF	
			,	$,	INS
aber	but		aber	KON	
			eigentlich	ADV	MOVT
sind	are		sind	VAFIN	
eigentlich	actually				MOVS
sie	they		sie	PPER	
nicht	not		nicht	PTKNEG	
praxisorientiert	practice-oriented		praxisorientiert	ADJD	

Target hypotheses

L2 (L1)		POS	TH	TH POS	DIFF
Mnn	[man|one]	PIS	Man	PIS	CHA
muss	must	VMFIN	muss	VMFIN	
sich	oneself				DEL
mit	with	APPR	mit	APPR	
diesen	these	PDAT	diesen	PDAT	
Theorien	theories	NN	Theorien	NN	
umgehen	deal	VVINF	umgehen	VVINF	
können	can	VMINF	können	VMINF	
			,	$,	INS
aber	but	KON	aber	KON	
			eigentlich	ADV	MOVT
sind	are	VAFIN	sind	VAFIN	
eigentlich	actually	ADV			MOVS
sie	they	PPER	sie	PPER	
nicht	not	PTKNEG	nicht	PTKNEG	
praxisorientiert	practice-oriented	ADJD	praxisorientiert	ADJD	

Target hypotheses

L2 (L1)		POS	TH	TH POS	DIFF
Mnn	[man|one]	PIS	Man	PIS	CHA
muss	must	VMFIN	muss	VMFIN	
sich	oneself				DEL
mit	with	APPR	mit	APPR	
diesen	these	PDAT	diesen	PDAT	
Theorien	theories	NN	Theorien	NN	
umgehen	deal	VVINF	umgehen	VVINF	
können	can	VMINF	können	VMINF	
			,	$,	INS
aber	but	KON	aber	KON	
			eigentlich	ADV	MOVT
sind	are	VAFIN	sind	VAFIN	
eigentlich	actually	ADV			MOVS
sie	they	PPER	sie	PPER	
nicht	not	PTKNEG	nicht	PTKNEG	
praxisorientiert	practice-oriented	ADJD	praxisorientiert	ADJD	

Outline

Related work I – syntactic analysis of learner data

- Only few studies on learner data looking beyond lexical data:
  - Menzel & Schröder (1999) developed an experimental system for automatic analysis of learner language in the context of diagnosis in tutoring systems
  - Dickinson & Ragheb (2009) describe a dependency-based annotation scheme for learner language
  - Rosén and de Smedt (2010) discuss strategies for syntactic analysis of learner data and argue for a semi-automatic approach based on a treebank of corrected second language (L2) texts, complemented with error annotations of the original L2 data
  - Meurers et al. (2010) work at creating a longitudinal learner corpus of reading comprehension questions; Ott and Ziai (2010) manually annotated parts of the reading comprehension corpus with dependency structure
- Until now there exists no syntactically annotated corpus of learner language for German (and not many for other languages)

Related work I – syntactic analysis of learner data

- Only few studies on learner data looking beyond lexical data:
  - Menzel & Schröder (1999) developed an experimental system for automatic analysis of learner language in the context of diagnosis in tutoring systems
  - Dickinson & Ragheb (2009) describe a dependency-based annotation scheme for learner language
  - Rosén and de Smedt (2010) discuss strategies for syntactic analysis of learner data and argue for a semi-automatic approach based on a treebank of corrected second language (L2) texts, complemented with error annotations of the original L2 data
  - Meurers et al. (2010) work at creating a longitudinal learner corpus of reading comprehension questions; Ott and Ziai (2010) manually annotated parts of the reading comprehension corpus with dependency structure
- Until now there exists no syntactically annotated corpus of learner language for German (and not many for other languages)

Related work II – impact of POS tags on parsing

- Quality of POS tags has high impact on parsing accuracy
  - Reported decrease in parsing results (f-score) for automatically predicted POS tags in the range of
    - 0.6-1.8% on German **newspaper text** (Petrov & Klein, 2008)
    - 2-3% on the same data (Rafferty & Manning, 2008)
- Accuracy of POS tagging of English as a second language is substantially lower than for native language (Haan, 2000; van Rooy and Schäfer, 2003; Meunier & Mönnink, 2001)
- POS accuracy decreases when applying the tagger to a new domain (Coden et al, 2005; Miller et al., 2006; Kübler & Baucom, 2011)

We expect a strong effect for L2 / new domain data on POS tagging/parsing accuracy

Related work II – impact of POS tags on parsing

- Quality of POS tags has high impact on parsing accuracy
  - Reported decrease in parsing results (f-score) for automatically predicted POS tags in the range of
    - 0.6-1.8% on German **newspaper text** (Petrov & Klein, 2008)
    - 2-3% on the same data (Rafferty & Manning, 2008)
- Accuracy of POS tagging of English as a second language is substantially lower than for native language (Haan, 2000; van Rooy and Schäfer, 2003; Meunier & Mönnink, 2001)
- POS accuracy decreases when applying the tagger to a new domain (Coden et al, 2005; Miller et al., 2006; Kübler & Baucom, 2011)

We expect a strong effect for L2 / new domain data on POS tagging/parsing accuracy

Related work II – impact of POS tags on parsing

- Quality of POS tags has high impact on parsing accuracy
  - Reported decrease in parsing results (f-score) for automatically predicted POS tags in the range of
    - 0.6-1.8% on German **newspaper text** (Petrov & Klein, 2008)
    - 2-3% on the same data (Rafferty & Manning, 2008)
- Accuracy of POS tagging of English as a second language is substantially lower than for native language (Haan, 2000; van Rooy and Schäfer, 2003; Meunier & Mönnink, 2001)
- POS accuracy decreases when applying the tagger to a new domain (Coden et al, 2005; Miller et al., 2006; Kübler & Baucom, 2011)

We expect a strong effect for L2 / new domain data on POS tagging/parsing accuracy

Related work II – impact of POS tags on parsing

- Quality of POS tags has high impact on parsing accuracy
  - Reported decrease in parsing results (f-score) for automatically predicted POS tags in the range of
    - 0.6-1.8% on German **newspaper text** (Petrov & Klein, 2008)
    - 2-3% on the same data (Rafferty & Manning, 2008)
- Accuracy of POS tagging of English as a second language is substantially lower than for native language (Haan, 2000; van Rooy and Schäfer, 2003; Meunier & Mönnink, 2001)
- POS accuracy decreases when applying the tagger to a new domain (Coden et al, 2005; Miller et al., 2006; Kübler & Baucom, 2011)

We expect a strong effect for L2 / new domain data on POS tagging/parsing accuracy

Parsing learner data

- Our data
  - non-canonical/highly marked structures
  - new domain (argumentative essays)
- Idea: support the parser by providing gold POS tags
  - keep effort for manual correction low:
    compare different strategies for manual correction
  - record time requirements and impact on parsing results

Outline

FALKO

- FALKO – **F**ehler-**A**nnotiertes **L**erner**KO**rpus
  (error-annotated learner corpus)

  (Lüdeling et al. 2008, Reznicek et al. 2010)

  - argumentative essays (4 topics)
  - by advanced learners (university students): **124.524** tokens
  - control corpus:
    essays by German L1 highschool/university students:
    68.940 tokens

- Target hypotheses (TH) for L2 and L1 data

POS tag correction

- **Assumption**:
  POS quality has high impact on parsing accuracy
- **Idea**:
  Improve parsing quality by semi-automatic correction of POS
- **Questions:**
  Is it enough to correct only some of the POS tags?
  - use different taggers to predict POS
  - correct only those tags where taggers disagree
  - correct only those tags where taggers disagree and at least one tagger predicted a verb
- Time requirements / impact on parsing?

Experimental setup

- **Tagger:**
  - TreeTagger (Schmid, 2004)
  - RFTagger (Schmid & Laws, 2008)
  - Stanford POS tagger (Toutanova et al., 2003)
- **Tag set**: STTS (Schiller et al., 1995)
- **Data**: Falko TH for L2 (248 essays) and L1 (94 essays)

	description	**no. sentences**
FALKO	test set for assessing tagger quality	125
	coder training set	594
	batches 1 - 12	6000
	FALKO200 gold standard	200
TiGer	parser training set	48.474

Experimental setup II

- **Gold standard: FALKO200**
  - 200 sentences randomly extracted from FALKO (L1: 100 sent., L2: 100 sent.)
  - manual correction of automatically predicted parses (Berkeley parser; Petrov & Klein, 2007)
  - each sentence corrected independantly by 2 annotators (5 post-graduate annotators with linguistic training)

- **Pilot study**
  - How many errors do we ignore when only correcting POS where taggers disagree?
  - 125 sentences L2, annotated from scratch
  - IAA on those sentences: 0.978 (Fleiss' κ)

tagger	acc.	no. err.
Stanford	0.962 %	72
TreeTagger	0.969 %	60
RFTagger	0.983 %	33
errors missed:	0.001 %	(2/1921 tokens)

Experimental setup II

- **Gold standard: FALKO200**
  - 200 sentences randomly extracted from FALKO (L1: 100 sent., L2: 100 sent.)
  - manual correction of automatically predicted parses (Berkeley parser; Petrov & Klein, 2007)
  - each sentence corrected independantly by 2 annotators (5 post-graduate annotators with linguistic training)

- **Pilot study**
  - How many errors do we ignore when only correcting POS where taggers disagree?
  - 125 sentences L2, annotated from scratch
  - IAA on those sentences: 0.978 (Fleiss' κ)

tagger	**acc.**	**no. err.**
Stanford	0.962 %	72
TreeTagger	0.969 %	60
RFTagger	0.983 %	33
errors missed:	0.001 %	(2/1921 tokens)

Time requirements for POS correction

batch	setting	# sent	# token corrected	time total avg.	time per tag avg.	time per tag coder1	time per tag coder2
1,2,5	*correct-all*	1500	1884	11198.02	6.25	6.16	6.35
3,4,6	*verb-only*	1500	587	3242.61	5.56	5.84	5.28

- substantial time savings for verb-only setting

Impact on parsing accuracy (FALKO200)

	L1				L2			
	prec	**rec**	**f-sc.**	**tag acc**	**prec**	**rec**	**f-sc.**	**tag acc**
	tagger-assigned POS tags							
stanf.	73.5***	74.0***	73.8	97.2	75.3***	77.1***	76.2	96.4
tree	75.5**	75.4**	75.4	98.0	76.2***	77.3***	76.7	97.8
rf	77.1 .	76.7	76.9	98.8	79.6	80.6	80.1	98.9
	parser-assigned POS tags							
berkley	**77.9**	**77.6**	**77.8**	98.2	80.0	80.6	80.3	97.7
	manually corrected POS tags							
A1(vo)	77.4	76.9	77.1	99.2	**80.5**	**81.0**	**80.8**	99.4
A2(vo)	77.8	77.5	77.7	99.9	80.4	**81.0**	80.7	99.9
A1(all)	77.5	76.9	77.2	99.3	80.1	80.7	80.4	99.3
A2(all)	77.4	77.1	77.2	99.6	79.7	80.6	80.1	99.6
gold	**77.9**	77.5	77.7	100.0	80.3	80.9	80.6	100.0

POS error correction – Results

- Despite same (TreeTagger) or higher tag acc. (RFTagger): parser benefits more when using its own POS

	L2	
	f-score	**tag acc**
TreeTagger	76.7	97.8
RFTagger	80.1	**98.9**
Berkeley	**80.3**	97.7

→ POS accuracy is not enough to predict parsing accuracy

Outline

Conclusions

- Semi-automatic POS correction as one step on the way towards a treebank of learner data
- Lessons learned:
  - THs are crucial for syntactic analysis of learner language

	L2 orig.	**L2 TH**
tag acc	93.8%	98.7%

  - no significant improvements of parsing accuracy on manually corrected POS
- **Outlook:** explore the adequacy of dependency representations for analysing learner language

Thank You!

Questions?

References

- Coden, Anni R., Serguei V. Pakhomov, Rie K. Ando, Patrick H. Duffy and Christopher G. Chute. 2005. Domain-specific language models and lexicons for tagging. Journal of Biomedical Informatics 38(6):422–430.
- Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards interlanguage pos annotation for effective learner corpora in SLA and FLT. Language Forum 36(1–2):139–154.
- Dickinson, Markus and Marwa Ragheb. 2009. Dependency annotation for learner corpora. In Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT-8), pages 59–70. Milan, Italy.
- Haan, Pieter de. 2000. Tagging non-native english with the TOSCA-ICLE tagger. In C. Mair, ed., Corpus linguistics and linguistic theory: Papers from the twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999 , vol. 33 of Language and computers, pages 69–79. Amsterdam: Rodopi.
- Kübler, Sandra, Eric Baucom: Fast Domain Adaptation for Part of Speech Tagging for Dialogues. RANLP 2011: 41–48.
- Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das Lernerkorpus Falko. Deutsch als Fremdsprache 2:67–73.
- Meunier, Fanny and Inge de Mönnink. 2001. Assessing the success rate of EFL learner corpus tagging: Online abstract. In ICAME 2001 Future Challenges in Corpus Linguistics. http://cecl.fltr.ucl.ac.be/Events/icamepr.htm.
- Menzel, Wolfgang and Ingo Schröder. Error diagnosis for language learning systems. ReCALL, (special edition, May 1999):20-30, 1999.

References

- Meurers, Detmar, Niels Ott, and Ramon Ziai: "Creation and Analysis of a Reading Comprehension Exercise Corpus: Towards Evaluating Meaning in Context". Accepted for publication in: Thomas Schmidt and Kai Wörner, Multilingual Corpora and Multilingual Corpus Analysis. Hamburg Studies in Multilingualism (HSM). Benjamins.
- Miller, John E., Michael Bloodgood, Manabu Torii, and K. Vijay-Shanker. 2006. Rapid adaptation of POS tagging for domain specific uses. In Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology (LNLBioNLP '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 118–119.
- Ott, Nils and Ramon Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. In: Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9). Tartu, Estonia, 3â4 December, 2010.
- Improved Inference for Unlexicalized Parsing, Slav Petrov and Dan Klein, In proceedings of HLT-NAACL 2007.
- Petrov, Slav and Dan Klein. 2008. Parsing German with latent variable grammars. In Proceedings of the Workshop on Parsing German, PaGeâ08, pages 33â39. Columbus, Ohio.
- Rafferty, Anna N. and Christopher D. Manning. 2008. Parsing three German treebanks: lexicalized and unlexicalized baselines. In Proceedings of the Workshop on Parsing German, PaGeâ08, pages 40â46. Columbus, Ohio.
- Reznicek, Marc, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas. 2010. Das Falko-Handbuch: Korpusaufbau und Annotationen. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin.

References

- Rosén, Victoria and Koenraad De Smedt. 2010. Syntactic annotation of learner corpora. In H. Johansen, A. Golden, J. E. Hagen, and A.-K. Helland, eds., Systematisk, variert, men ikke tilfeldig, pages 120–132. Novus forlag.
- van Rooy, B., & Schäfer, L. (2003). An evaluation of three POS taggers for the tagging of the Tswana learner English corpus. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 conference, 28-31 March 2003 (vol. 16 of University Centre For Computer Corpus Research On Language Technical Papers) (pp. 835-844). Lancaster, UK: Lancaster University.
- Schiller A., Teufel S., Stöckert C. and Thielen C. (1999) Guidelines für das Tagging deutscher Textcorpora, University of Stuttgart / University of Tübingen, also available at `www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html`
- Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
- Schmid, Helmut, Florian Laws (2008): Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, COLING 2008, Manchester, Great Britain.
- Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.