Alignment Weighting for Short Answer Assessment

Björn Rudzewitz

brzdwtz@sfs.uni-tuebingen.de

University of Tübingen

February 17, 2016

Motivation

- Highly variable language learner data
- Can computers diagnose the meaning of learner answers ?
- Robustness vs. Recall

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Introduction

Alignment Weighting

- **Goal**: automatic assessment of semantic correctness short answers by language learners
- **Context**: reading comprehension in L2
  - learners read a text
  - write free-text answers to question to the text
  - system predicts whether answer is correct
- **Problem**: answers are very diverse
- **Challenge**: design a system that is
  - robust enough for high variability
  - doesn't gloss over important aspects
- **Solution**: define/use importance of answer aspects

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Data

- Corpus of Reading Comprehension Exercises in German (CREG) [Meurers et al., 2010]
- reading texts, questions, learner answers, target answers
- high diversity and variability in interlanguage
- longitudinal corpus collection $\rightarrow$ different proficiency levels represented
- high variability on surface and semantic level
- labels: form-independent answer correctness

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Data

- dimension of diversity: **question types**
- free text answers
- some questions allowing more freedom in answer [Meurers et al., 2011]
- Example: *why*/*who* vs. *Yes*/*No* question

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Data

Example

Question:
Wann wurde der Euro im täglichen Gebrauch in den Ländern der EU eingeführt ?

Reference Answer:
Am 1. Januar 2002 wurde der Euro im täglichen Gebrauch in den Ländern der EU eingeführt.

Learner Answer:
Eine gemeinsame Wahrung, der Euro, wurde am 1. Januar 1999.

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

System

- Comparing Meaning in Context system (CoMiC) [Meurers et al., 2011]
- alignment-based short answer assessment system
- goal: predict whether a learner answer is semantically (in-)correct, **without** attention to form errors
- 3 steps:
  1. Annotation
  2. Alignment
  3. Diagnosis

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

System in a nutshell

- **Annotation**:
  sentences, tokens, lemmas, pos tags, dependencies, synonyms, spelling correction, pmi-similarity, chunks
- **Alignment**:
  - Traditional Marriage Algorithm [Gale and Shapley, 1962] $\rightarrow$ globally optimal alignment configuration between student and target answer
  - result: 1 token aligned to 0 or 1 other tokens on a specific level of annotation
- **Diagnosis**
  - features to measure number and kinds of alignments
  - machine learning component for predictions

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Alignment Example

Question: *Was kann man in Dresden noch machen ?*

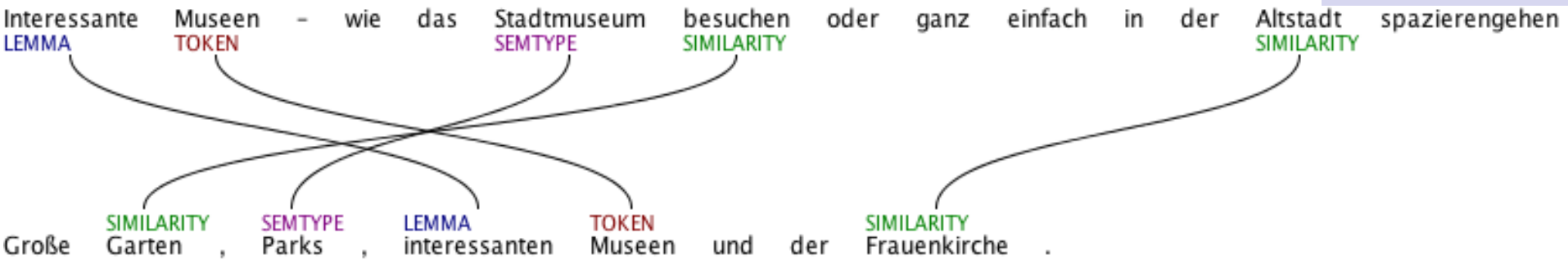

Figure: Alignment between target answer (top) and student answer (bottom) on different levels.

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Alignment Weighting

- baseline system counts number and kinds of alignments
- however: context determines which elements are more important for a meaning diagnosis
- some questions may support higher answer diversity/variability
- Example: in a *who* question, it may be more important to align (proper) nouns
- **Idea**: weight alignments by their relative importance

Alignment Weighting

- important factors for answer diagnosis:
  - task context
  - task language
- 3 alignment weightings on these dimensions:
  - task-based weighting
  - L2-based weighting
  - hybrid weighting: task + L2 weighting

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Task Weighting

- task context important for producing and assessing answers
- **operationalization**: question types
- add new features to the system to encode question type
- **idea**: learn question-type specific alignment patterns and variation in diversity

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

L2 Weighting

- general properties of L2 important for producing and assessing answers
- **operationalization**: part of speech (POS) tags
- **idea**: a certain question type may require alignments of words of a certain syntactic category
- encoding of (un)symmetricity of POS tag alignment distributions over equivalence classes
- STTS main word classes [Schiller et al., 1995]
- different perspectives, different normalizations

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Hybrid Weighting

- **idea**: measure how important aligned elements are in corresponding reading text
- term frequency inverse document frequency (*tf.idf*) measure: how important are words in one document in comparsion to other documents
- feature for encoding *tf.idf* values of aligned elements

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Experimental Testing

- extrinsic evaluation: test whether the CoMiC system performs more accurately with alignment features
- do alignment weighting features help with linguistic diversity ?
- test single alignment weightings and combinations
- interactions: model question-type specific part of speech alignment patterns and see how important aligned elements are

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Experimental Testing

Results in a nutshell

features	CREG-1032-KU	CREG-1032-OSU
CoMiC	84.6	87.0
CoMiC + POS	85.2	90.0*
CoMiC + tf.idf	86.1	88.4
CoMiC + q-types	85.4	87.2
CoMiC + all	87.9*	86.5

Table: Accuracy of the system with different weightings. * denotes a statistically significant improvement (McNemar's test, $\alpha = 0.05$)

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Capturing Diversity

Question-Type Specific Accuracy

q-type	# inst.	Align. Weight.	Meurers et al. [2011]
Alternative	7	0.57	0.57
How	144	**0.91**	0.86
What	276	**0.87**	0.86
When	6	**1.00**	0.86
Where	9	0.67	**88.9**
Which	170	**0.93**	0.92
Why	174	**0.84**	0.79
Who	41	**0.85**	0.83
Yes/No	5	0.80	1.00
Several	200	**0.83**	0.77
Total	1032	**87.0**	84.6

Table: Macro-averages of the CoMiC system with alignment weighting on CREG-1032 obtained by grouping results by question type. Boldface indicates an improvement over the contrastive approach.

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Results

Observations

- alignment weighting improves accuracy
- *tf.idf* weighting, POS weighting always effective
- question types alone not highly effective, but in interaction
- improvement for many question types
- especially for questions with high answer diversity (*why*/*who*)

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Discussion

- Ziai and Meurers [2014]: CoMiC + information structure: 90.3%
- Horbach et al. [2013]: CoMiC-reimplementation + pos-align criteria + use of reading text: 84.4%
- Hahn and Meurers [2012]: CoSeC (semantic for alignment): 86.3%
- Pado and Kiefer [2015]: answer clustering, threshold diagnosis: 83.7%

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

Conclusions

- novel approach to encode importance of alignments
- highly beneficial for short answer assessment, significant improvements
- problem of assessing higher linguistic diversity for some question types partly overcome

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

References

David Gale and Lloyd S Shapley. College Admissions and the Stability of Marriage. *American Mathematical Monthly*, pages 9–15, 1962.

Michael Hahn and Detmar Meurers. Evaluating the Meaning of Answers to Reading Comprehension Questions A Semantics-Based Approach. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 326–336. Association for Computational Linguistics, 2012.

Andrea Horbach, Alexis Palmer, and Manfred Pinkal. Using the text to evaluate short answers for reading comprehension exercises. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 1, pages 286–295, 2013.

Detmar Meurers, Niels Ott, Ramon Ziai, et al. Compiling a Task-Based Corpus for the Analysis of Learner Language in Context. *Proceedings of Linguistic Evidence. Tübingen*, pages 214–217, 2010.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9. Association for Computational Linguistics, 2011.

Ulrike Pado and Cornelia Kiefer. Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA*, page 43, 2015.

Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. *Manuscript, Universities of Stuttgart and Tübingen*, 66, 1995.

Ramon Ziai and Detmar Meurers. Focus Annotation in Reading Comprehension Data. *LAW VIII*, page 159, 2014.

Björn Rudzewitz

brzdwtz@sfs.uni-tuebingen.de

University of Tübingen

Appendix: q-type pos align patterns

q-type	#inst.	10 most informative Part of Speech tags
Alternative	7	VVPP, PPOSAT, PPER, PPOS, VMFIN, PRELAT, PIS, PIDAT, PIAT, PDS
How	144	NN, CARD, VVFIN, ADJA, ART, VAFIN, NE, PIAT, PRELS, PTKNEG
What	276	NN, KON, ADJA, VVPP, VVINF, APPRART, PIS, CARD, PTKNEG, PWAV
When	6	ADV, KOKOM, KOUS, NN, PIS, PWF, PIDAT, PWAV, PPOSAT, VAFIN
Where	9	PIDAT, PPER, PPOSAT, PRELAT, PIS, VVPP, PRF, PIAT, PAVDAT
Which	170	NN, ADV, VVPP, PTKNEG, VAFIN, NE, VAINF, CARD, KON, PIS
Why	174	NN, VVFIN, ART, APPR, PIAT, VAFIN, KON, NE, ADJA, KOKOM
Who	41	NN, VVINF, ADJD, VMFIN, PPER, PRELAT, PRELS, PPOS, PPOSAT, PTKANT
Yes/No	5	PTKANT, PPOSAT, PRELAT, PPOS, PIS, PPER, PIDAT, PRF, PIAT, PAV
Several	200	NN, NE, ADJA, PIAT, VMFIN, KON, PIS, VVPP, KON, PTKNEG

Table: Most informative part of speech alignments by question type.

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen

POS Tag Equivalence Classes

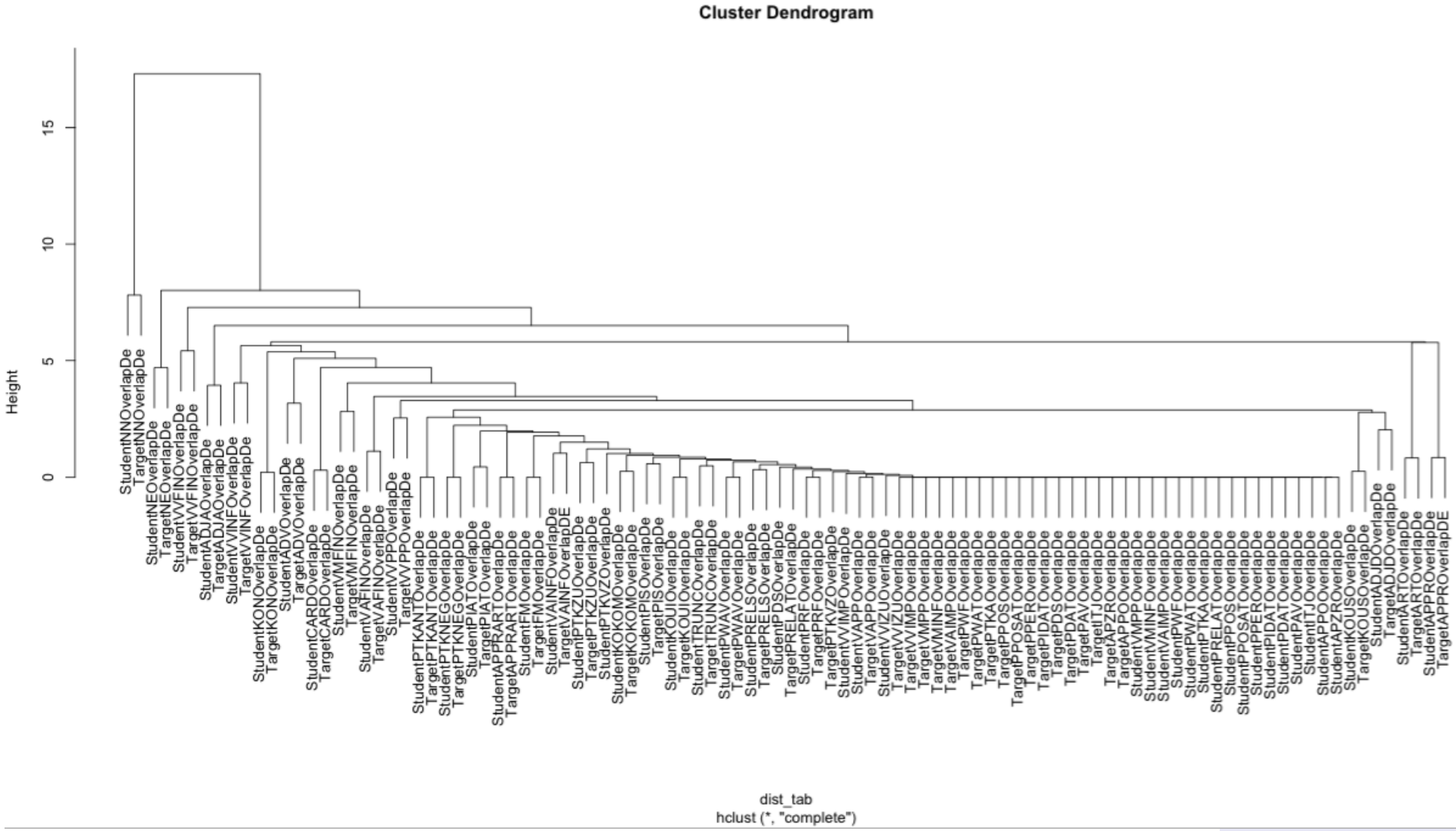

Figure: Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.

POS Tag Equivalence Classes

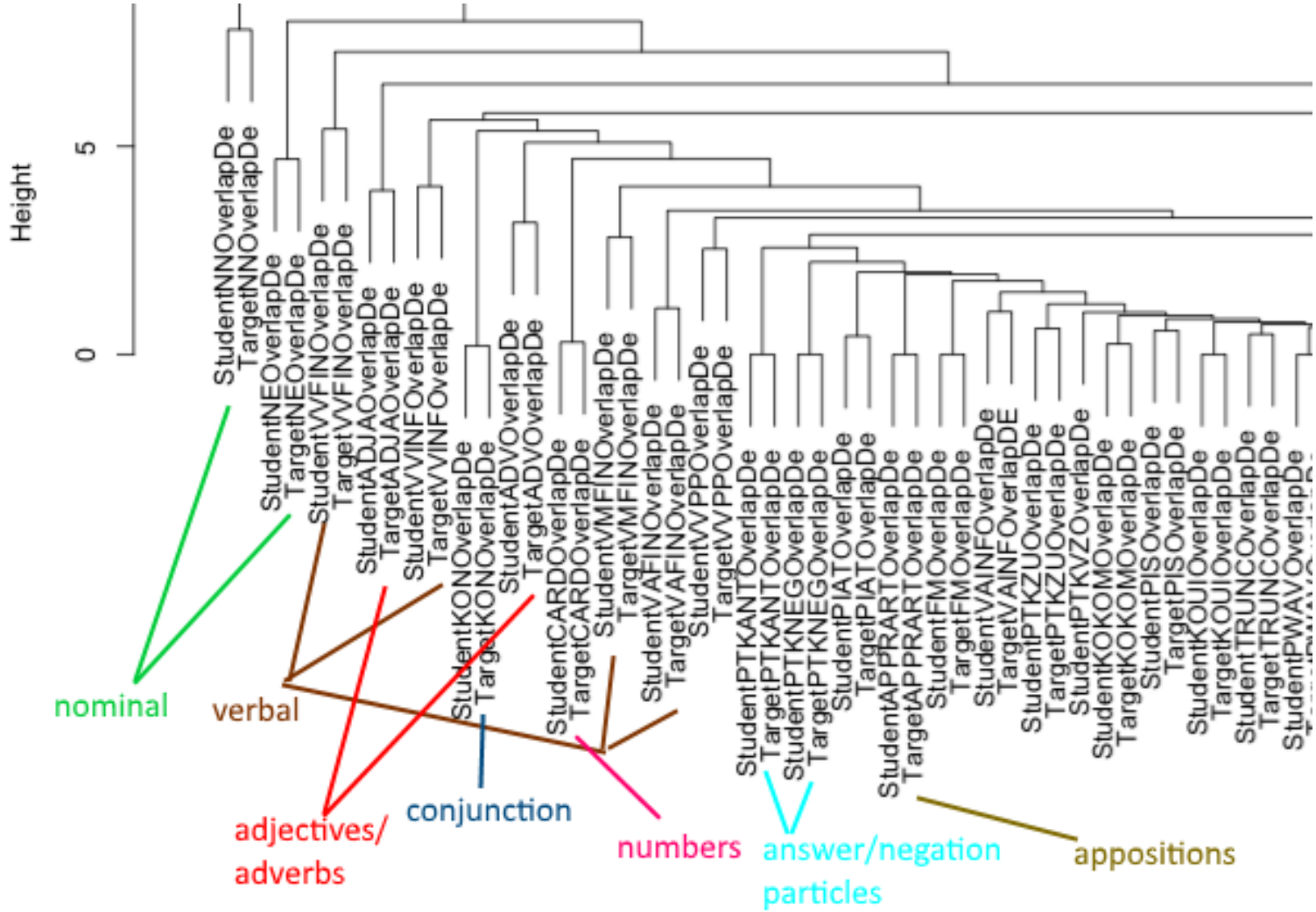

Figure: Part of Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.

Björn Rudzewitz
brzdwtz@sfs.uni-tuebingen.de
University of Tübingen