# Alignment Weighting for Short Answer Assessment

Björn Rudzewitz[1]
University of Tübingen

Presentation of B.A. Thesis

October 30, 2015

[1] bjoern.rudzewitz@student.uni-tuebingen.de

# Reading Comprehension

Reading comprehension in foreign language learning context:

- text
- questions
- target answers
- student (language learner) answers

# Reading Comprehension

Learners need to ...

- ... understand the text and questions
- ... use L2 to formulate answers

# Reading Comprehension

Learners need to ...

- ... understand the text and questions

  → **task** competence

- ... use L2 to formulate answers

  → **language** competence / performance

# Reading Comprehension

Learners need to ...

- ... understand the text and questions

  → **task** competence

- ... use L2 to formulate answers

  → **language** competence / performance

Goal of this work: incorporate aspects of concrete task and general language in automatic SAA approach by alignment weighting

# Data : CREG

Corpus of Reading Exercises in German [Meurers et al., 2010]

- longitudinal learner corpus collected at 2 German programs in USA (OSU, KU)
- structure:
  - texts
  - questions
  - target answers (TA)
  - student answers (SA)
  - meta data
  - links between elements (SA $\rightarrow$ TA, SA $\rightarrow$ Diagnosis,...)
- significant variation / deviation of form and meaning in SAs
- binary (and detailed) gold diagnosis of *semantic* correctness of SAs

# Data: CREG

Various subsets used for experiments

| data set | # questions | # SAs | # TAs |
|---|---|---|---|
| CREG-1032-KU | 117 | 610 | 180 |
| CREG-1032-OSU | 60 | 422 | 147 |
| CREG-3620-KU | 89 | 735 | 181 |
| CREG-3620-OSU | 585 | 2885 | 705 |
| CREG-5K-KU | 214 | 1814 | 382 |
| CREG-5K-OSU | 663 | 3324 | 875 |

Table: Data distribution of CREG subsets used in this study.

# Baseline System

**CoMiC-DE** system [Meurers et al., 2011]

- Comparing Meaning in Context
- alignment-based short answer assessment system
- UIMA pipeline [Ferrucci and Lally, 2004]
- goal: diagnose form-independent meaning of SAs

# CoMiC: System Architecture

3-step approach:

1. *Annotation*
   use NLP tools to generate linguistic multi-layer markup
2. *Alignment*
   use annotations to align similar elements between SA and TA
3. *Diagnosis*
   use features measuring quantity and quality of alignments for binary diagnosis

# CoMiC: System Architecture

3-step approach:

1. *Annotation*
   use NLP tools to generate linguistic multi-layer markup
2. *Alignment*
   use annotations to align similar elements between SA and TA
3. *Diagnosis*
   use features measuring quantity and quality of alignments for binary diagnosis

# CoMiC Phase 1: Annotation

| Task | NLP Tool |
|---|---|
| Sentence Detection | OpenNLP[Baldridge, 2005] |
| Tokenization | OpenNLP [Baldridge, 2005] |
| Lemmatization | TreeTagger [Schmid, 1994] |
| Spell Checking | Edit distance [Levenshtein, 1966] , igerman98 word list |
| Part of Speech Tagging | TreeTagger [Schmid, 1994] |
| Noun Phrase Chunking | OpenNLP [Baldridge, 2005] |
| Lexical Relations | GermaNet [Hamp et al., 1997] |
| Similarity Score | PMI-IR [Turney, 2001] |
| Dependency Relations | MaltParser [Nivre et al., 2007] |

Table: NLP tools used in the CoMiC-DE system.

# CoMiC: System Architecture

3-step approach:

1. *Annotation*
   use NLP tools to generate linguistic multi-layer markup
2. *Alignment*
   use annotations to align similar elements between SA and TA
3. *Diagnosis*
   use features measuring quantity and quality of alignments for binary diagnosis

# CoMiC Phase 2: Alignment

- align tokens, chunks, dependency triples
- elements given in question are excluded
- alignment candidates: words with overlaps on various linguistic levels
- use TMA [Gale and Shapley, 1962] for annotation matching
- alignment annotation contains alignment label

# CoMiC Phase 2: Alignment

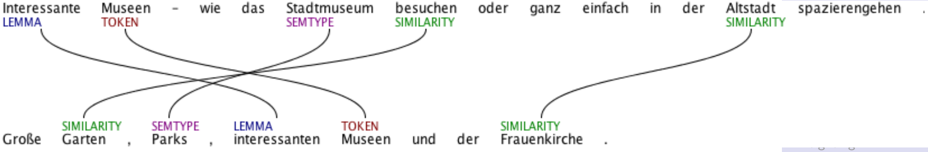


Figure: Alignment between target answer (top) and student answer (bottom) on different levels.



Björn Rudzewitz
University of Tübingen

# CoMiC: System Architecture

3-step approach:

1. *Annotation*
   use NLP tools to generate linguistic multi-layer markup
2. *Alignment*
   use annotations to align similar elements between SA and TA
3. *Diagnosis*
   use features measuring quantity and quality of alignments for binary diagnosis

# CoMiC Phase 3: Diagnosis

- extract number and kinds of alignments for each SA
  $\rightarrow$ 13 ml features
- use TiMBL Daelemans et al. [2004] for LOO k-NN classification
- result: binary diagnosis for each SA

# CoMiC Phase 3: Diagnosis

| Feature | Description |
|---|---|
| 1. Keyword Overlap | % keywords aligned |
| 2. TA Token Overlap | % aligned TA tokens |
| 3. Learner Token Overlap | % aligned SA tokens |
| 4. TA Chunk Overlap | % aligned TA chunks |
| 5. Learner Chunk Overlap | % aligned SA chunks |
| 6. TA Triple Overlap | % aligned TA dependency triples |
| 7. Learner Triple Overlap | % aligned SA dependency triples |
| 8. Token Match | % token-identical token alignments |
| 9. Similarity Match | % similarity-resolved token alignments |
| 10. Type Match | % type-resolved token alignments |
| 11. Lemma Match | % lemma-resolved token alignments |
| 12. Synonym Match | % synonym-resolved token alignments |
| 13. Variety | Number of kinds of token-level alignments (features 8-12) |

Table: CoMiC baseline features.

# Alignment Weighting: Motivation

Idea:

- aligned elements have different properties
- alignments between certain elements may be more important

$\rightarrow$ weight existing alignments in new dimension of similarity

# Alignment Weighting

2 conceptual weighting approaches
$\rightarrow$ 3 implementations

1. General Linguistic Weighting
2. Task-Specific Weighting
3. Hybrid Approach

global vs. local weighting schemes

# General Linguistic Weighting

- weighting of aligned elements by language-wide property in new dimension of similarity
- operationalization of abstract concept of general linguistic property:
  **part of speech tag classes**
- pos tags represent syntactic, semantic, morphological language-wide properties



Björn Rudzewitz
University of Tübingen

# General Linguistic Weighting

- problem: data sparsity
- solution: abstraction/generalization via equivalence classes of outcomes
  $\rightarrow$ pos tag *classes*

How to find equivalence classes:

- *top-down* approach:
  using linguistic intuition to form classes of tags
- *bottom-up* approach:
  induce classes of tags from sample data

# Option 1: top-down approach

| Group | STTS tags |
|---|---|
| nominal | NN, NE |
| verbal | VVFIN, VVIMP, VVINF, VVIZU, VVPP, VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP |
| adjv | ADJA, ADJD, ADV |
| rest | APPR, APPRART, APPO, APZR, ART, CARD, FM, ITJ, KOUI, KOUS, KON, KOKOM, PDS, PDAT, PIS, PIAT, PIDAT, PPER, PPOSS, PPOSAT, PRELS, PRELAT, PRF, PWS, PWAT, PWAV, PAV, PTKZU, PTKNEG, PTKVZ, PTKANT, PTKA, TRUNC |

Table: Coarse STTS subsets used for the general linguistic weighting, adapted from [Rudzewitz and Ziai, 2015].

# Option 2: bottom-up approach

- choose a development set
- output single pos features for every tag for TA and SA
- perform hierarchical agglomerative clustering
- use clusters as equivalence classes for features

# Option 2: bottom-up approach

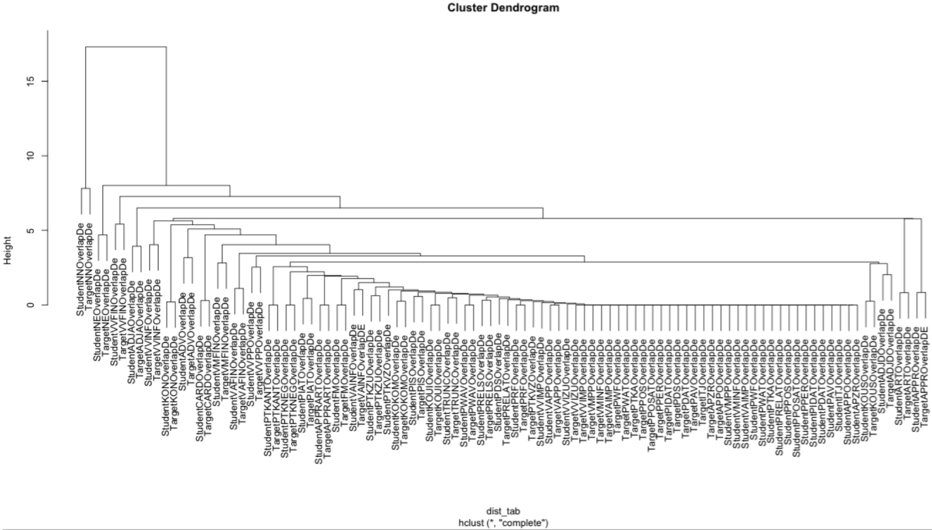


Figure: Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.

# Option 2: bottom-up approach

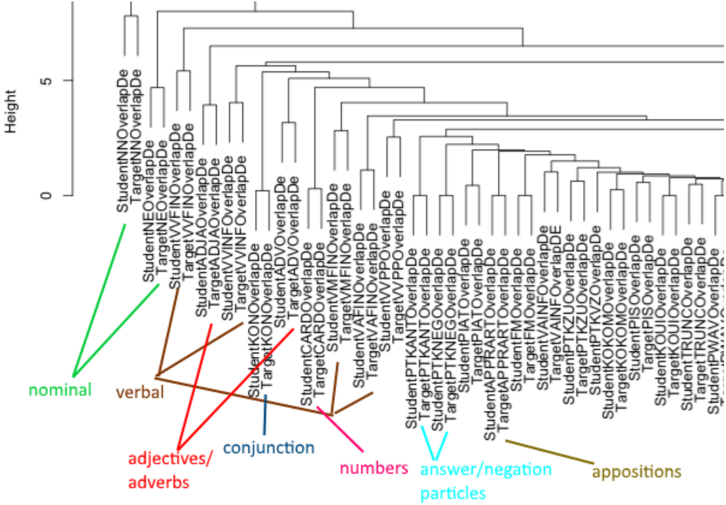


Figure: Part of Hierarchical Agglomerative Clustering of Part of Speech Tags over all instances of CREG-1032.

# Option 2: bottom-up approach

- observation: distinct clusters are representatives for 'main word' classes defined in STTS tag set [Schiller et al., 1995]
- hclust algorithm is given no assumptions about main word classes !

→ use STTS main word classes as equivalence classes

# Feature Variants

- problem with features: how to normalize ?
- more concrete: given numeric quantities of aligned elements, how to account for effects of answer length ?
- solution (in this work): explore and report results for all variants

# Feature Variants

$A_h \in A(\text{"Answers"}),\ w_j \in W_{A_h} \subset W(\text{"Words"}),\ t_{w_j} \in T_i \subset T(\text{"tag from tag group"})$

$$ol(A_h, T_i) = \frac{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [w_j \text{ is aligned AND } t_{w_j} = t \text{ AND } w_j \text{ is new}]}{\sum_{t \in T_i} \sum_{w_j \in W_{A_h}} [\text{see Table !}]}$$

| variant | $t_{w_j} = t$ | $w_j$ is new | $w_j$ is aligned |
|---|---|---|---|
| local | ✓ | ✓ | |
| semi-global | | ✓ | ✓ |
| global | | ✓ | |

Table: Denominator constraints for different feature variants. Logical conjunction AND between row values.

# Feature Variant Interpretation

- *local*: Are many of the new tokens with this part of speech tag aligned ?
- *semi-global*: Are many of the aligned tokens from a certain part of speech group ?
- *global*: Do many of the new words have a tag from this part of speech group and are at the same time aligned ?

# Interpolated Features

$$ol_{ip}(A_h, T_i) = ol_{local}(A_h, T_i) \times ol_{sglobal}(A_h, T_i) \times ol_{global}(A_h, T_i)$$

$$ol_{lip}(A_h, T_i) = \frac{1}{3} \times (ol_{local}(A_h, T_i) + ol_{sglobal}(A_h, T_i) + ol_{global}(A_h, T_i))$$

- combine the different feature variants

# Task-Specific Weighting

- goal: include the specific (local) task context in SAA
- "*task*": complex concept, many aspects
- operationalization: implement question-type features
- binary indicator function for each question type
- gold standard from previous study [Meurers et al., 2011] as development set
- 11 types: *Alternative*, *How*, *What*, *When*, *Where*, *Which*, *Who*, *Why*, *Yes/No*, *Several*, *Unknown*

# Hybrid Weighting Approach

- *tf.idf* lemma-based weighting, adapted from Manning and Schütze [1999]
- generally applicable measure, but task-specific training
- document collection: all reading texts in CREG-5K
- for each aligned token, get *tf.idf* weight in reading text to which the SA refers

$$ol_{tf.idf}(A_h) = \sum_{w_j \in W_{A_h}} weight_{tf.idf}(w_j, d_i)$$

$$weight_{tf.idf}(w_j, d_i) = \begin{cases} 0 & \text{, if } (w_j \text{ NOT new}) \text{ OR} \\ & (w_j \text{ NOT aligned}) \text{ OR} \\ & (w_j \notin d_i) \\ (1 + log(tf_{j,i})) \times log\frac{N}{df_j} & \text{, otherwise} \end{cases}$$

# Experimental Testing

Significance Testing: McNemar's test ($\alpha = 0.05$)

$H_0$: The binary classification performance of an alignment-based short answer assessment system does not change if it is augmented with part of speech or *tf.idf* features.

$H_1$: The binary classification performance of an alignment-based short answer assessment system significantly improves if it is augmented with part of speech or *tf.idf* features.

# Experimental Testing: Coarse POS

| system | 3620-KU | 3620-OSU | 1032-KU | 1032-OSU | 5K-KU | 5K-OSU |
|---|---|---|---|---|---|---|
| base | 81.5 | 82.2 | 84.6 | 87.0 | 80.9 | 82.5 |
| local | 82.0 | 82.6 | 85.2 | **90.0*** | 82.0 | 82.8 |
| semi-global | 81.2 | **84.1*** | 85.4 | 87.2 | 81.3 | **84.0*** |
| global | 83.0 | **83.6*** | 84.8 | 85.8 | 81.6 | **83.6*** |
| ip | 80.5 | **84.1*** | 85.1 | 85.1 | 81.7 | **84.4*** |
| lip | 82.6 | **84.1*** | 84.4 | 87.0 | 81.4 | **84.1*** |

Table: System performance for the baseline system augmented with part of speech features in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

# Experimental Results: Question Types and tf.idf

| system variant | 3620-KU | 3620-OSU | 1032-KU | 1032-OSU | 5K-KU | 5K-OSU |
|---|---|---|---|---|---|---|
| baseline | 81.5 | 82.2 | 84.6 | 87.0 | 80.9 | 82.5 |
| q-types | 80.8 | **83.1**$^*$ | 85.4 | 87.2 | 80.9 | 82.8 |

Table: System performance for the baseline system augmented with question type features in terms of accuracy. The symbol $^*$ denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

| system variant | 3620-KU | 3620-OSU | 1032-KU | 1032-OSU | 5K-KU | 5K-OSU |
|---|---|---|---|---|---|---|
| baseline | 81.5 | 82.2 | 84.6 | 87.0 | 80.9 | 82.5 |
| tf.idf | **84.2**$^*$ | **84.1**$^*$ | 86.1 | 88.4 | **83.1**$^*$ | **84.3**$^*$ |

Table: System performance for the baseline system augmented with *tf.idf* features in terms of accuracy. The symbol $^*$ denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

# Experimental Testing: Combination

| system variant | 3620-KU | 3620-OSU | 1032-KU | 1032-OSU | 5K-KU | 5K-OSU |
|---|---|---|---|---|---|---|
| baseline | 81.5 | 82.2 | 84.6 | 87.0 | 80.9 | 82.5 |
| q-types + stts local + tf.idf | 83.8 | **84.7***| **87.9*** | 86.5 | 82.4 | 84.9 |
| q-types + stts semi-global+ tf.idf | 83.1 | **84.6*** | 85.4 | 88.2 | 82.1 | 84.9 |
| q-types + stts global+ tf.idf | **84.2*** | **84.5*** | **87.9*** | 84.6 | **82.6*** | **84.6*** |
| q-types + stts ip+ tf.idf | 83.3 | **84.7*** | **88.9*** | 84.1 | **82.8*** | **85.3*** |
| q-types + stts lip+ tf.idf | **84.5*** | **85.0*** | **88.0*** | 85.8 | **82.8*** | **84.9*** |

Table: System performance for the baseline system augmented with question type and STTS group part of speech features and *tf.idf* weighting in terms of accuracy. The symbol * denotes a statistically significant improvement over the baseline ($\alpha = 0.05$).

# Experimental Testing: Main results

- *many* more tables with accuracies and test statistics ...
- pos features alone result in highest accuracy on one data set (90%)
- *tf.idf* always yields improvement
- question-types alone not as effective
- best overall result for combination of all 3 weightings
- linguistically interpretable question-type specific pos alignment patterns (Appendix 1)
- question-type specific macro-averages show improvement from Meurers et al. [2011] (Appendix 2)

# Discussion: Related work

- Ziai and Meurers [2014]: CoMiC + information structure
- Horbach et al. [2013]: CoMiC-reimplementation + pos-align criteria + use of reading text
- Hahn and Meurers [2012]: CoSeC
- many other SAA systems (see thesis)

# Conclusion

- significant improvements with novel techniques
- results highly competitive to state-of-the-art systems
- no human annotation needed
- linguistically interesting insights from ml algorithms
- combination of all feature variants most effective

# Appendix 1: 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.

# Appendix 2: q-type macro-averages

| q-type | # inst. | local | sglobal | global | ip | lip |
|---|---|---|---|---|---|---|
| Alternative | 7 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 |
| How | 144 | **0.88** | **0.89** | **0.91** | **0.90** | **0.90** |
| What | 276 | **0.87** | **0.88** | **0.87** | 0.85 | **0.88** |
| When | 6 | **1.00** | 0.83 | **1.00** | 0.83 | 0.83 |
| Where | 9 | 0.67 | 0.56 | 0.67 | 0.67 | 0.67 |
| Which | 170 | 0.91 | 0.92 | **0.93** | 0.92 | 0.92 |
| Why | 174 | **0.84** | **0.84** | **0.84** | **0.83** | **0.84** |
| Who | 41 | **0.88** | **0.90** | **0.85** | **0.88** | **0.85** |
| Yes/No | 5 | 0.80 | 0.80 | 0.80 | 0.80 | 0.80 |
| Several | 200 | **0.86** | **0.83** | **0.83** | **0.86** | **0.85** |
| Micro | 1032 | **86.7** | **86.8** | **87.0** | **86.5** | **87.3** |

Table: Macro-averages of the best system variant on CREG-1032 obtained by grouping results by question type. Boldface indicates an improvement upon the results by Meurers et al. [2011]

Jason Baldridge. The OpenNLP Project. *URL: http://opennlp. apache. org/index. html,(accessed 25 August 2015)*, 2005.

Walter Daelemans, Jakub Zavrel, Kurt van der Sloot, and Antal Van den Bosch. TiMBL: Tilburg Memory-Based Learner. *Tilburg University*, 2004.

David Ferrucci and Adam Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

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.

Birgit Hamp, Helmut Feldweg, et al. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer, 1997.

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.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Christopher D Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

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.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135, 2007.

Björn Rudzewitz and Ramon Ziai. CoMiC: Adapting a Short Answer Assessment System for Answer Selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, volume 15, 2015.

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

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Citeseer, 1994.

Peter Turney. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. 2001.

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