KIT
Karlsruhe Institute of Technology

CSL
Cognitive Systems Lab

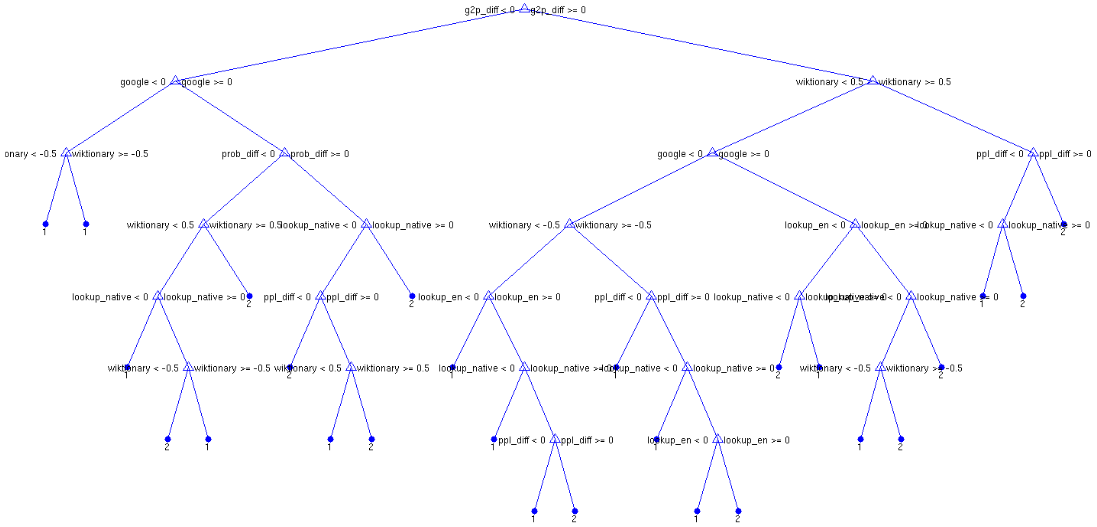
Automatic Detection of Anglicisms for the Pronunciation Dictionary Generation: A Case Study on our German IT Corpus

Sebastian leidig, **Tim Schlippe**, Tanja Schultz

SLTU 2014 – 4th Workshop on Spoken Language Technologies for Under-resourced Languages
St. Petersburg, Russia

*Zeigen
Sie
andere
Apps
für
einfaches
Multitasking
neben
dem
Browser
an
Internet
Explorer
nutzt
Hardwarebeschleunigung
Websites
werden
schneller
geladen
damit
Sie
noch
reibungsloser
surfen
können*

*Nimm
deine
Lieblingsmusik
überallhin
mit
kommt
der
iPod
shuffle
mit
Speicher
genug
für
hunderte
von
Songs
alle
wichtigen
Songs
fürs
Training
Wiedergabelisten
Genius
Mixes
Podcasts
und
Hörbücher*



Motivation

- From Microsoft's German website www.microsoft.de:
  - *“Zeigen Sie andere **Apps** für einfaches **Multitasking** neben dem **Browser** an.”*
  - *“**Internet Explorer** nutzt **Hardware**beschleunigung. **Websites** werden schneller geladen, damit Sie noch reibungsloser **surfen** können.”*

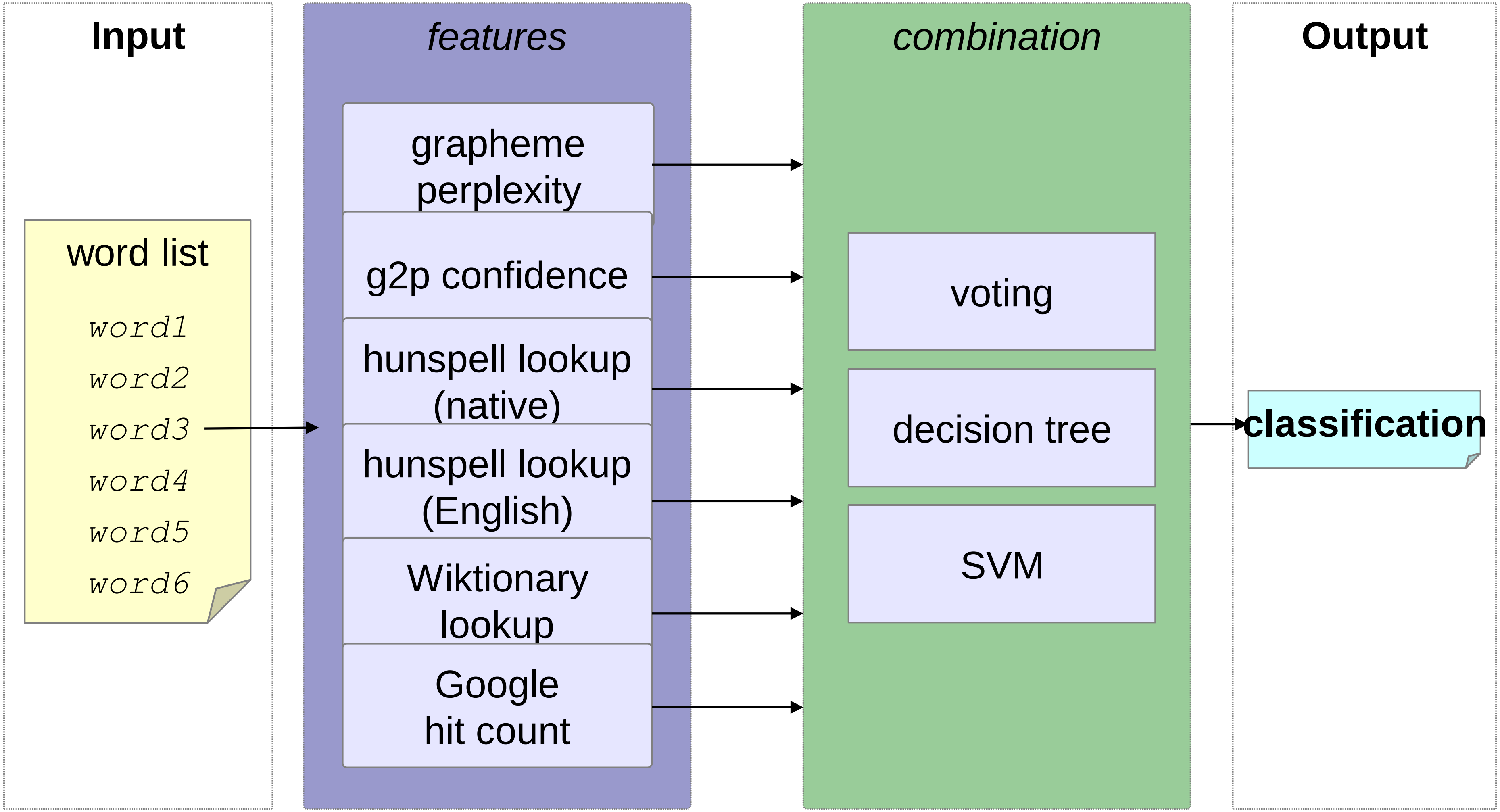
Motivation

- With the globalization words from other languages come into a language without assimilation to the phonetic system of the new language

- To economically build up lexical resources with automatic or semi-automatic methods

  → detect and treat them separately

Overview


Input
word list
word1
word2
word3
word4
word5
word6
features
grapheme perplexity
g2p confidence
hunspell lookup (native)
hunspell lookup (English)
Wiktionary lookup
Google hit count
combination
voting
decision tree
SVM
Output
classification

Outline

1. Motivation and Overview
2. Test Sets
3. Single Features
4. Combinations
5. Summary and Future Work

Test Sets - Domains

- **German IT website**
  - www.microsoft.de
  - 4.6k unique words
- **German general news**
  - www.spiegel.de
  - 6.6k unique words
- **Afrikaans**
  - NCHLT corpus (*Heerden, Davel, Barnard, 2013*), (*Basson, Davel, 2013*)
  - 9.4k unique words

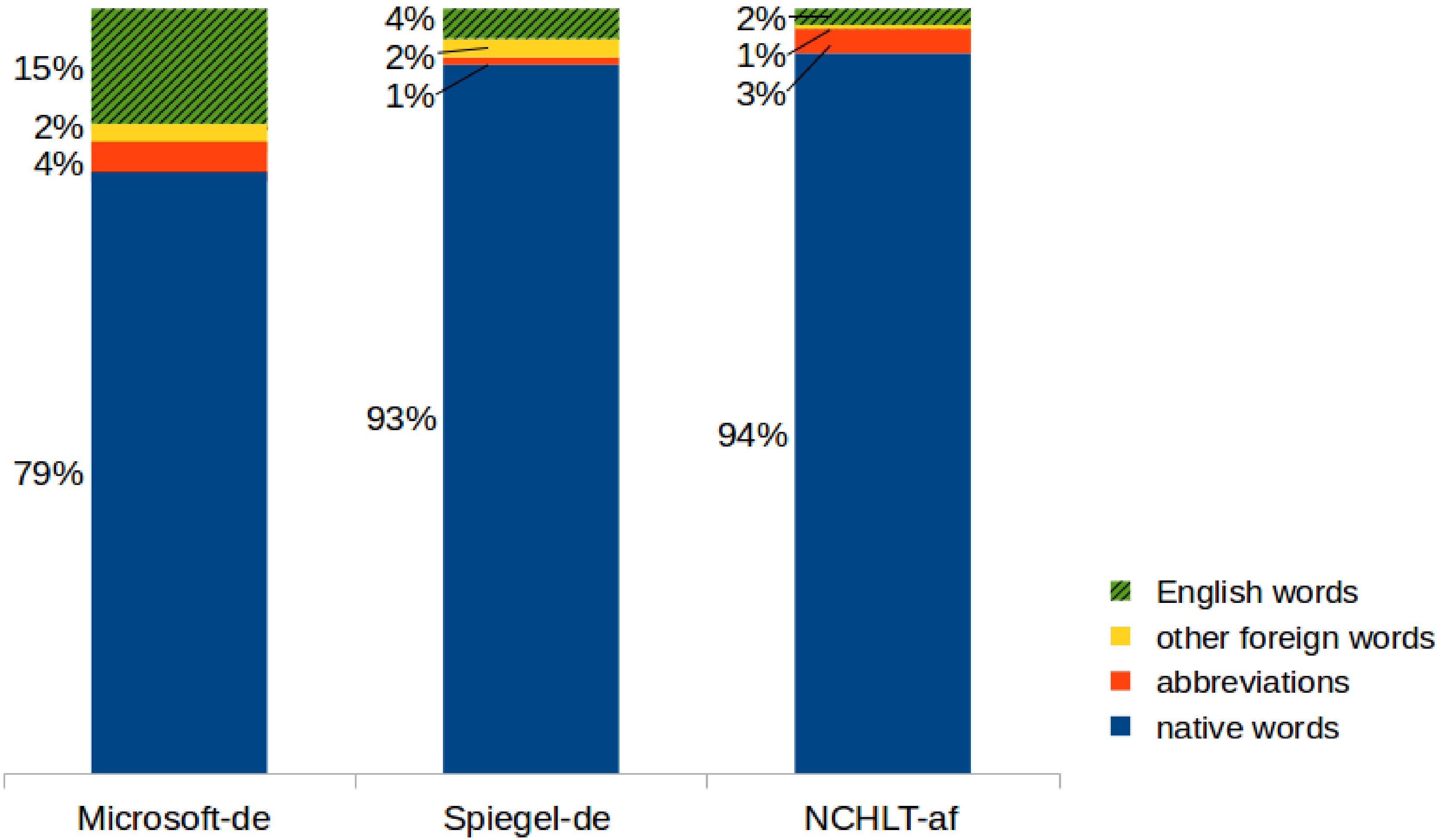
Test Sets - Domains

- Tag for
  - “**English**”:
    - e.g. Software, Brain, …
  - **Foreign hybrids**
    - Compound words
      - e.g. Schadsoftware, …
    - Grammatically adapted words
      - e.g. downloaden, …

- Different word categories:
  - **Abbreviations**:
    - e.g. UV, CIA, …
  - **Other foreign words**
    - Compound words
      - e.g. Français, Niveau, …

- Decisions based on
  - Agreement of annotators
  - duden.de

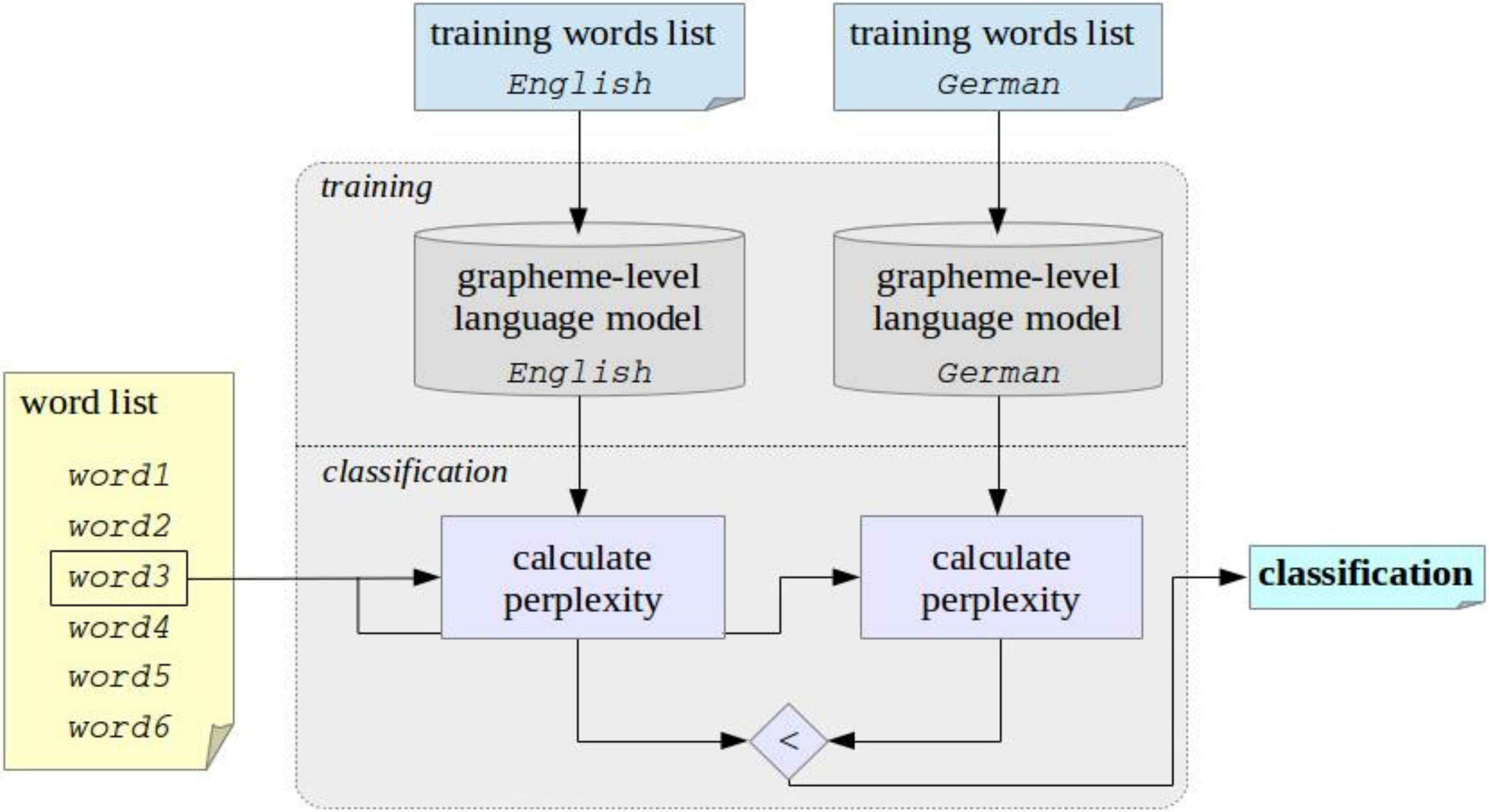
Foreign words in different test sets



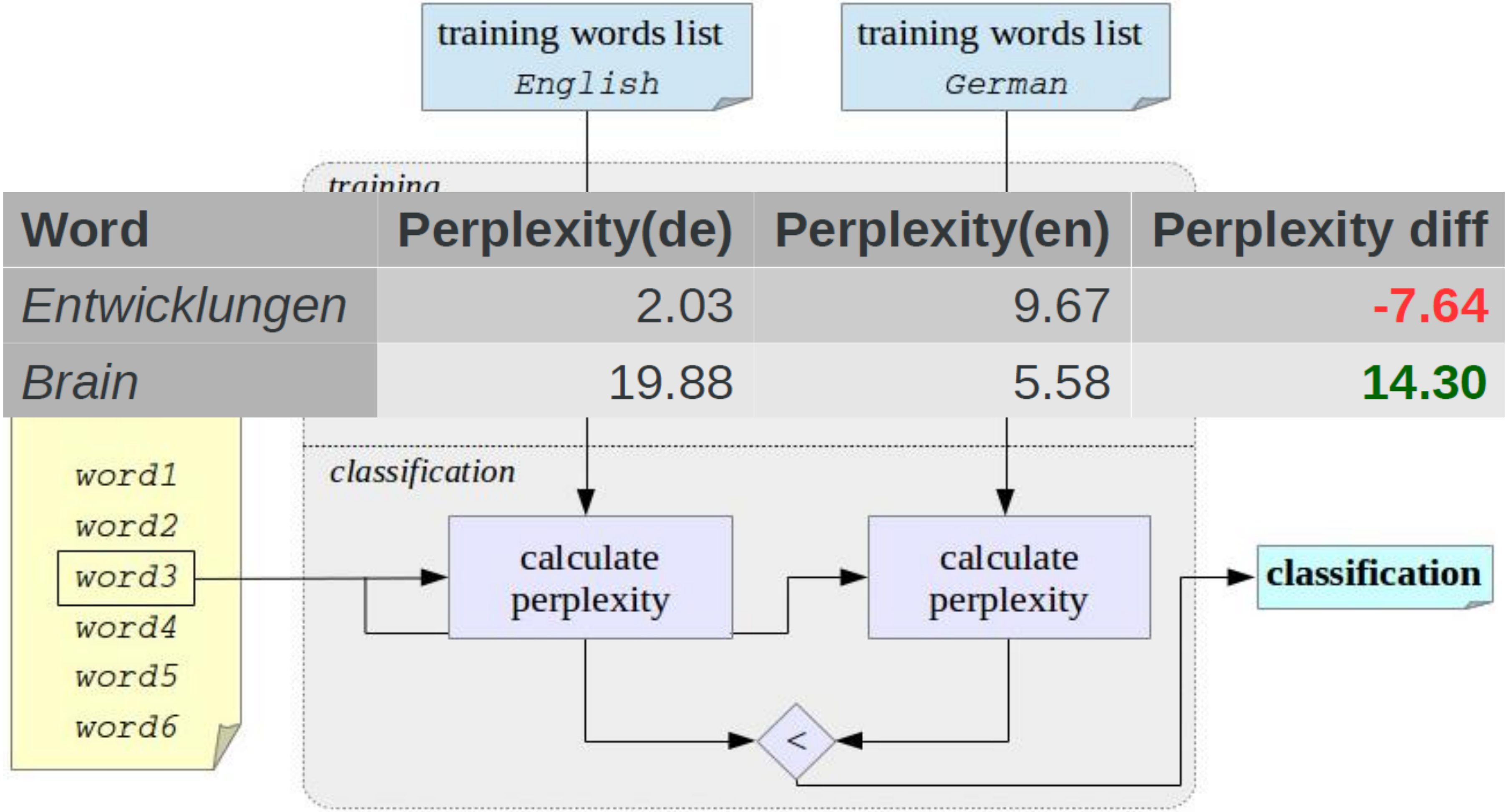
Single Features – Design Criteria

- Features trained on commonly available resources
  - Word lists, Pronunciation dictionaries, Spellchecker dictionaries, Wiktionary, Google
- Thresholds without supervised training
  - Comparison between English and native models
- New approaches

Grapheme Perplexity

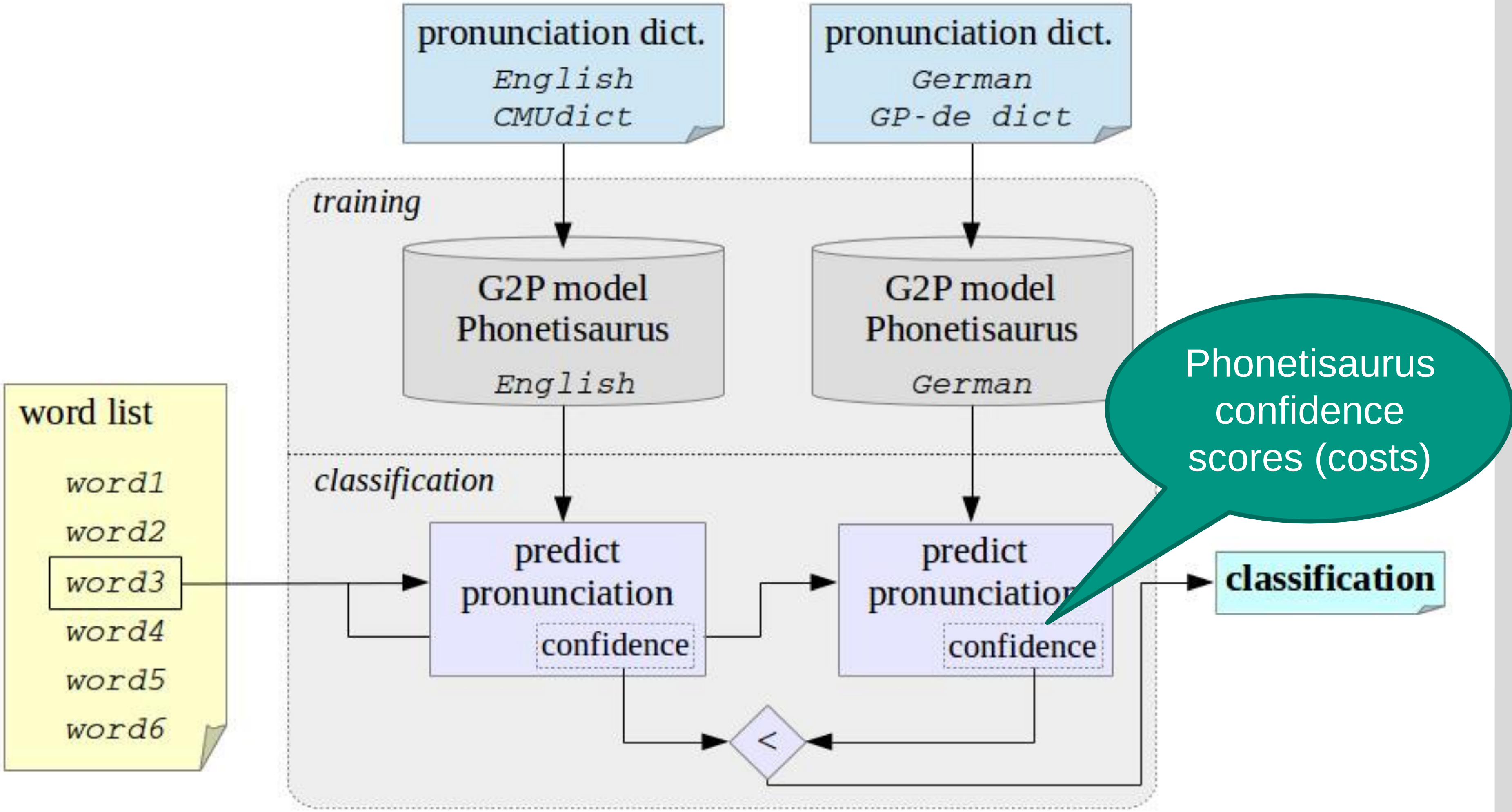


Grapheme Perplexity



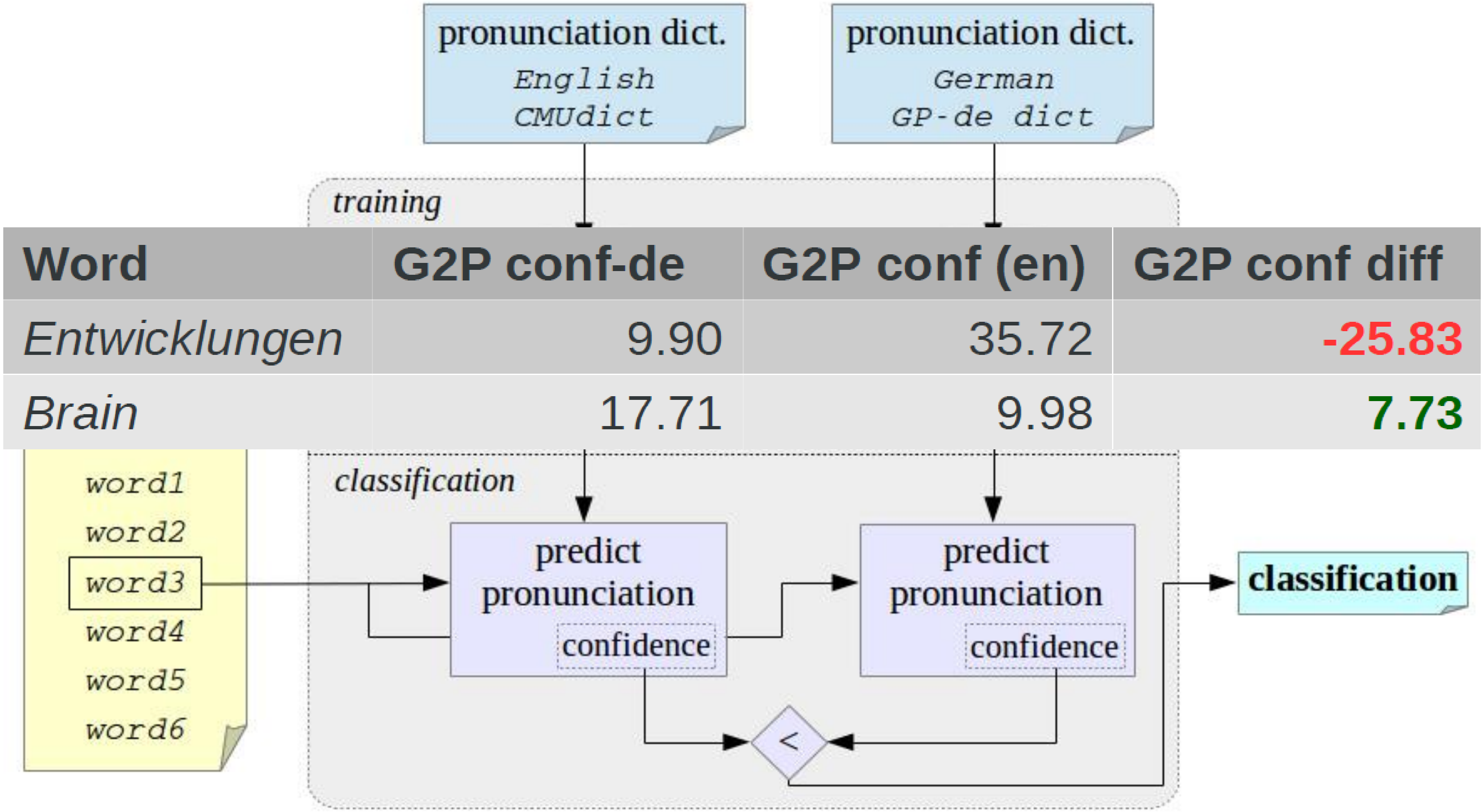
Word	Perplexity(de)	Perplexity(en)	Perplexity diff
Entwicklungen	2.03	9.67	-7.64
Brain	19.88	5.58	14.30

Grapheme-to-Phoneme Confidence



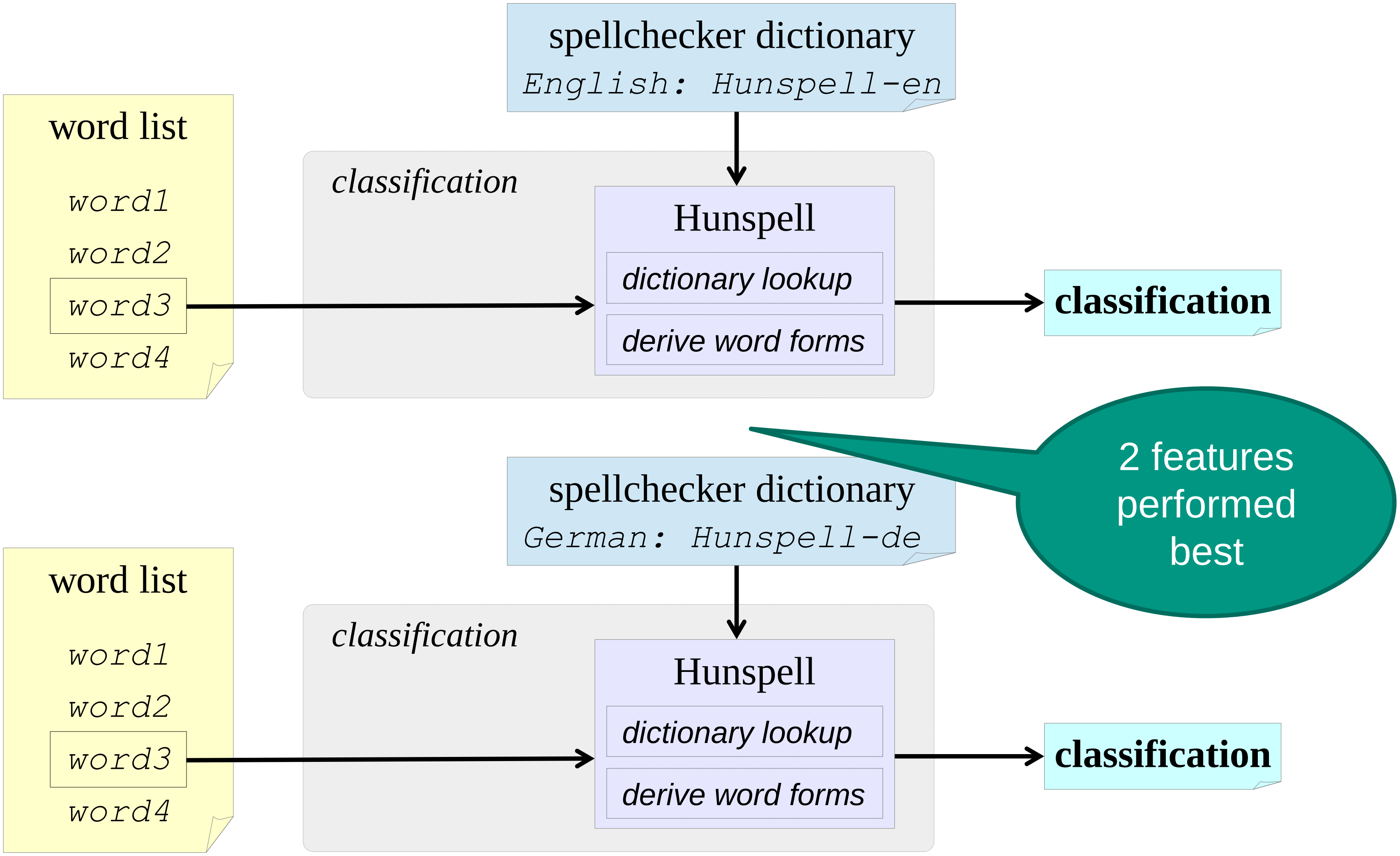
Phonetisaurus confidence scores (costs)

Grapheme-to-Phoneme Confidence



Word	G2P conf-de	G2P conf (en)	G2P conf diff
Entwicklungen	9.90	35.72	-25.83
Brain	17.71	9.98	7.73

Hunspell Lookup



spellchecker dictionary
English: Hunspell-en

word list
word1
word2
word3
word4

classification

Hunspell
dictionary lookup
derive word forms

classification

spellchecker dictionary
German: Hunspell-de

word list
word1
word2
word3
word4

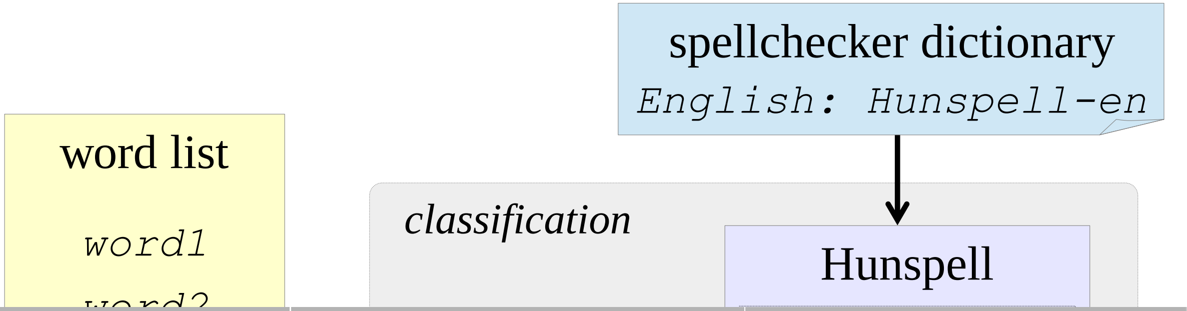
classification

Hunspell
dictionary lookup
derive word forms

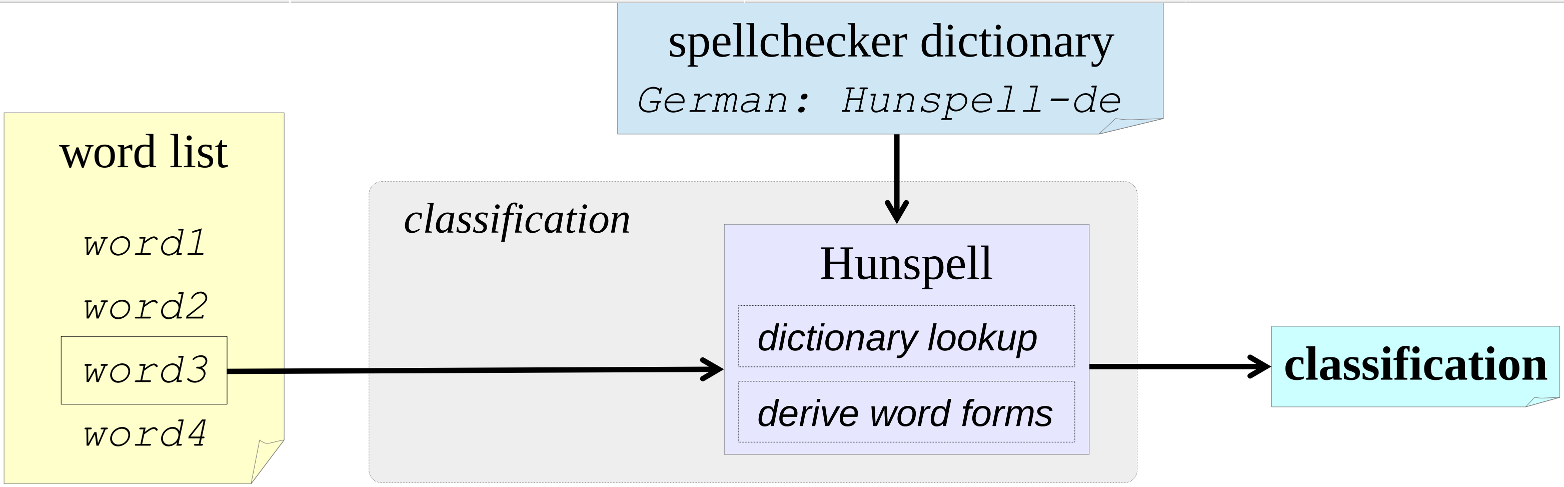
classification

2 features performed best

Hunspell Lookup

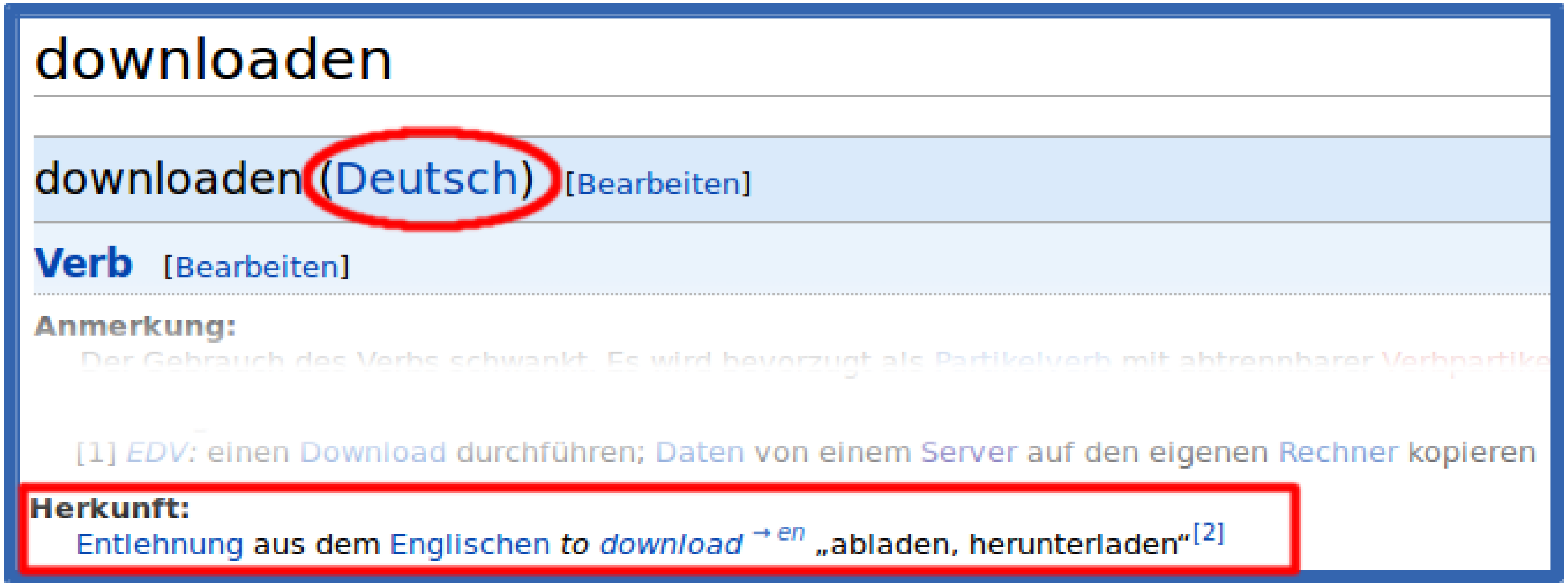


Word	Stem	Dictionary(de)	Dictionary(en)
Entwicklungen	→ Entwicklungs	yes	no
Brain	-	no	yes



Wiktionary Lookup

- Check crowdsourced information from matrix language Wiktionary



downloaden

downloaden (Deutsch) [Bearbeiten]

Verb [Bearbeiten]

Anmerkung:

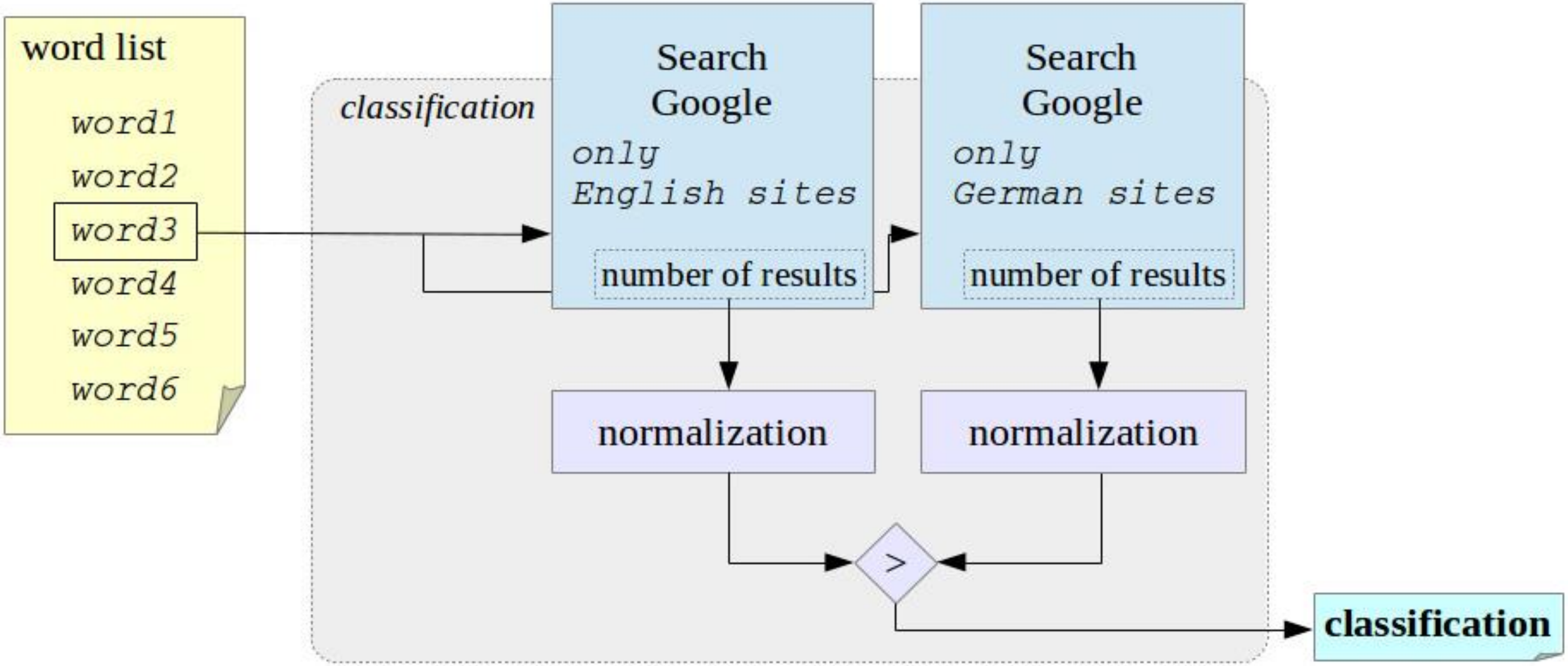
Der Gebrauch des Verbs schwankt. Es wird bevorzugt als Partikelverb mit abtrennbarer Verbpartike

[1] *EDV:* einen Download durchführen; Daten von einem Server auf den eigenen Rechner kopieren

Herkunft:

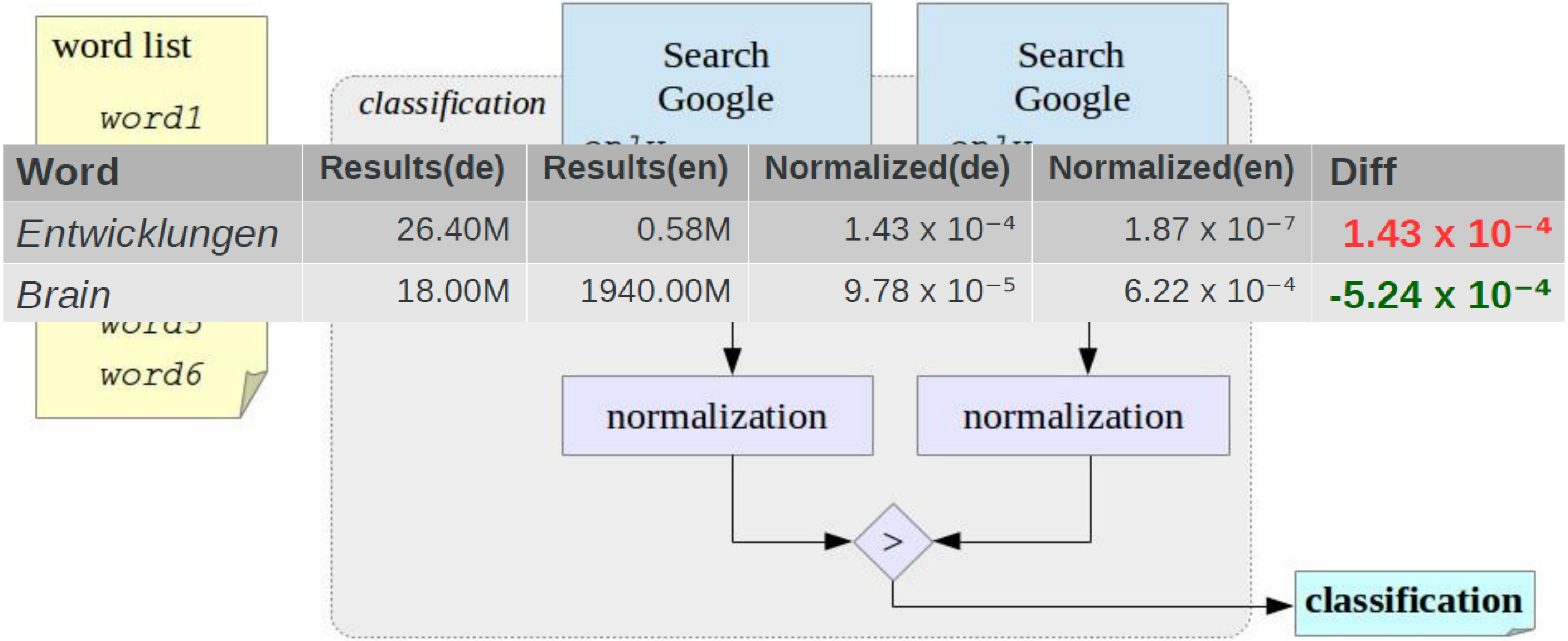
Entlehnung aus dem Englischen *to download* → en „abladen, herunterladen“[2]

Google Hit Count



- Based on Alex B. (2008) “Automatic Detection of English Inclusion in Mixed-lingual Data with an Application to Parsing”, University of Edinburgh

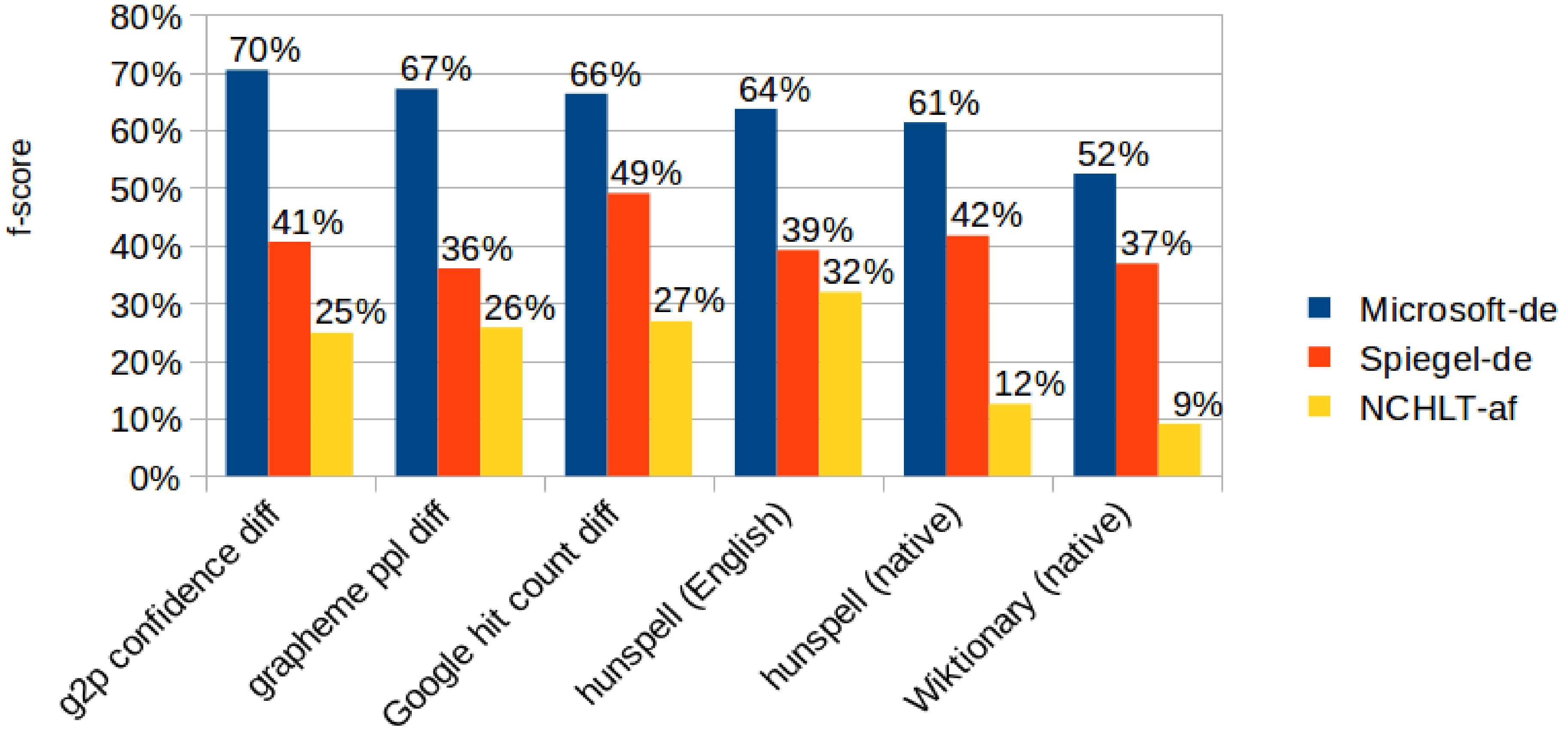
Google Hit Count



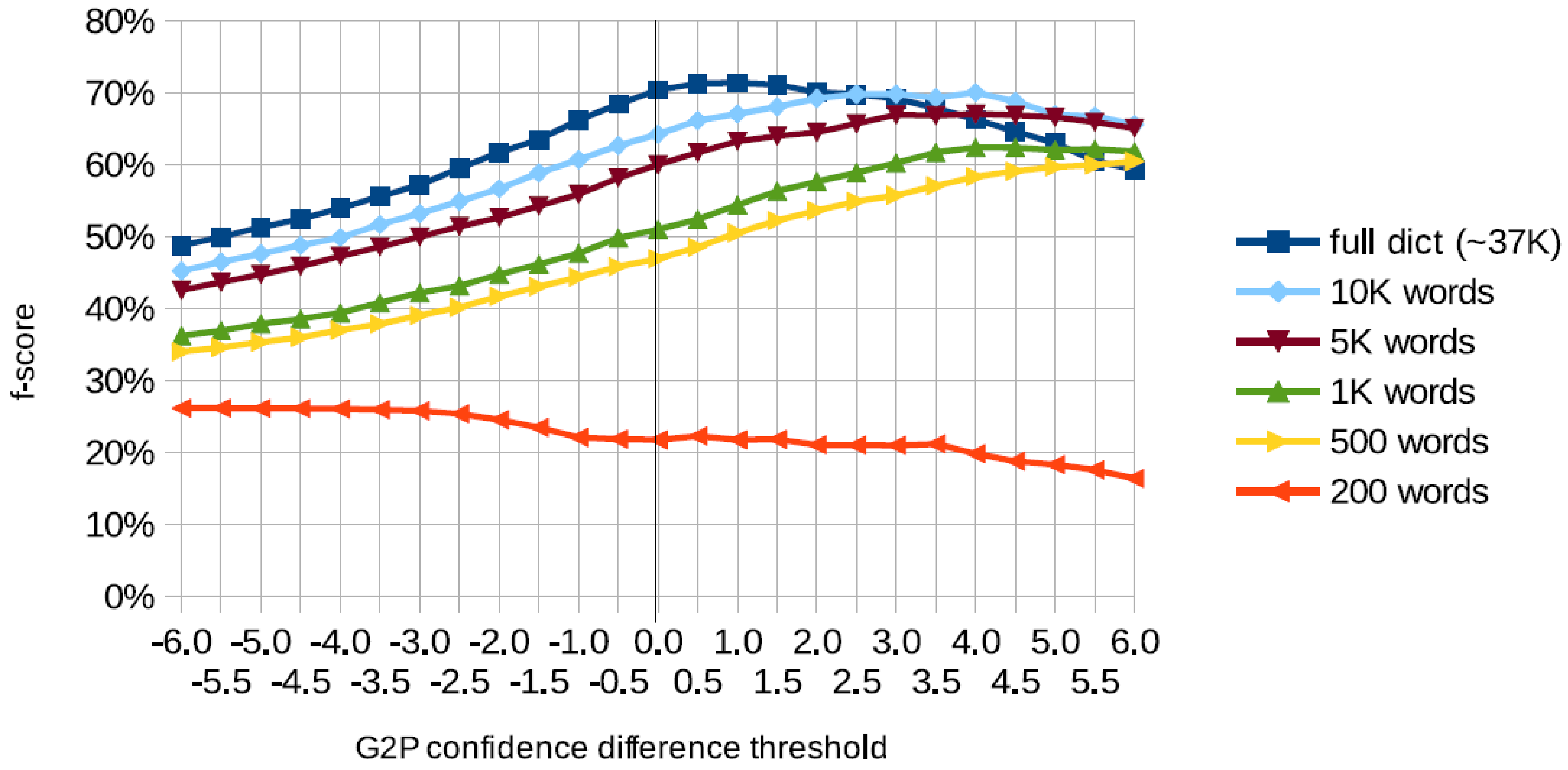
Word	Results(de)	Results(en)	Normalized(de)	Normalized(en)	Diff
Entwicklungen	26.40M	0.58M	1.43 x 10^{-4}	1.87 x 10^{-7}	**1.43 x 10^{-4}**
Brain	18.00M	1940.00M	9.78 x 10^{-5}	6.22 x 10^{-4}	**-5.24 x 10^{-4}**

- Based on Alex B. (2008) “Automatic Detection of English Inclusion in Mixed-lingual Data with an Application to Parsing”, University of Edinburgh

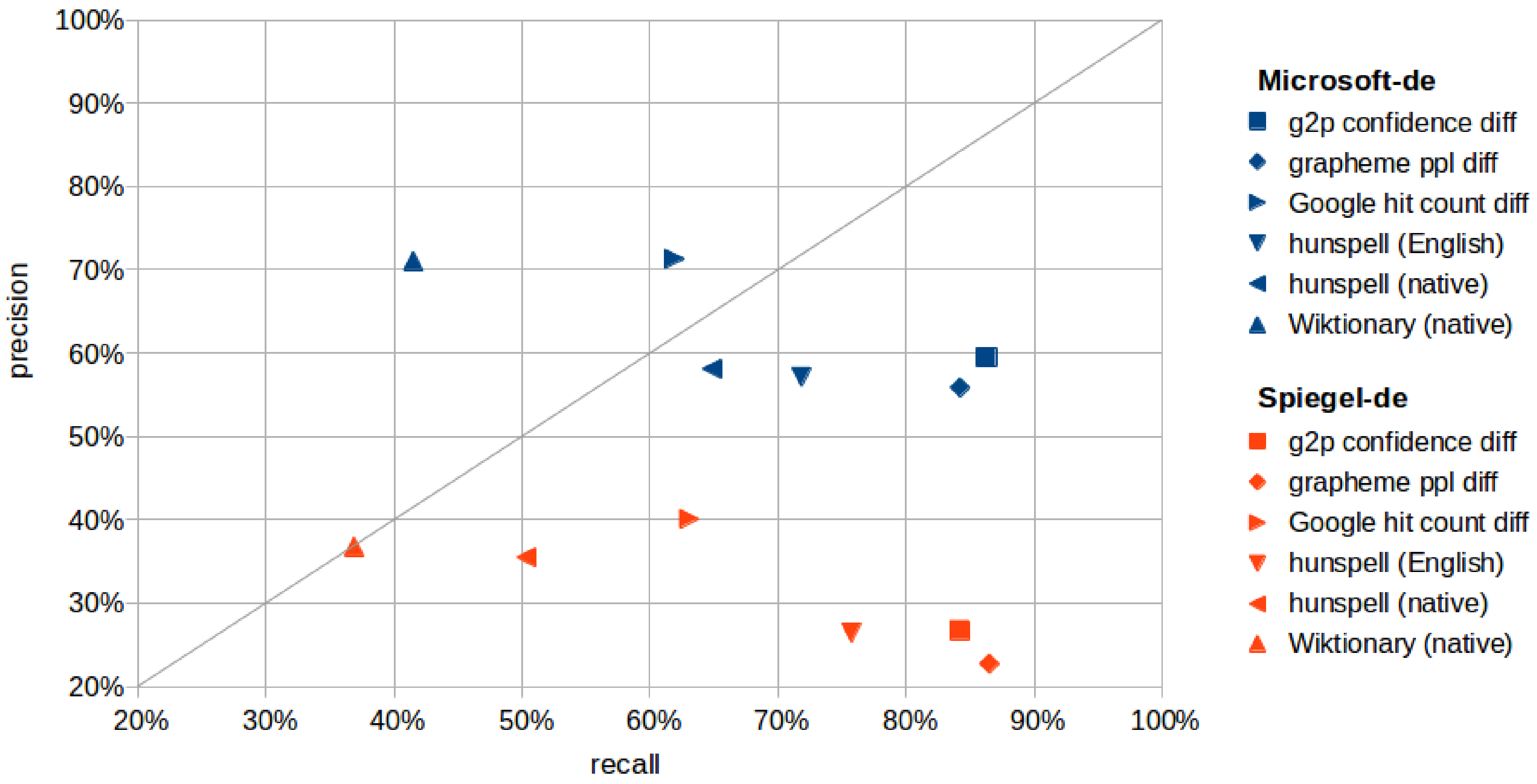
Result: Single Features



Grapheme-to-Phoneme Confidence

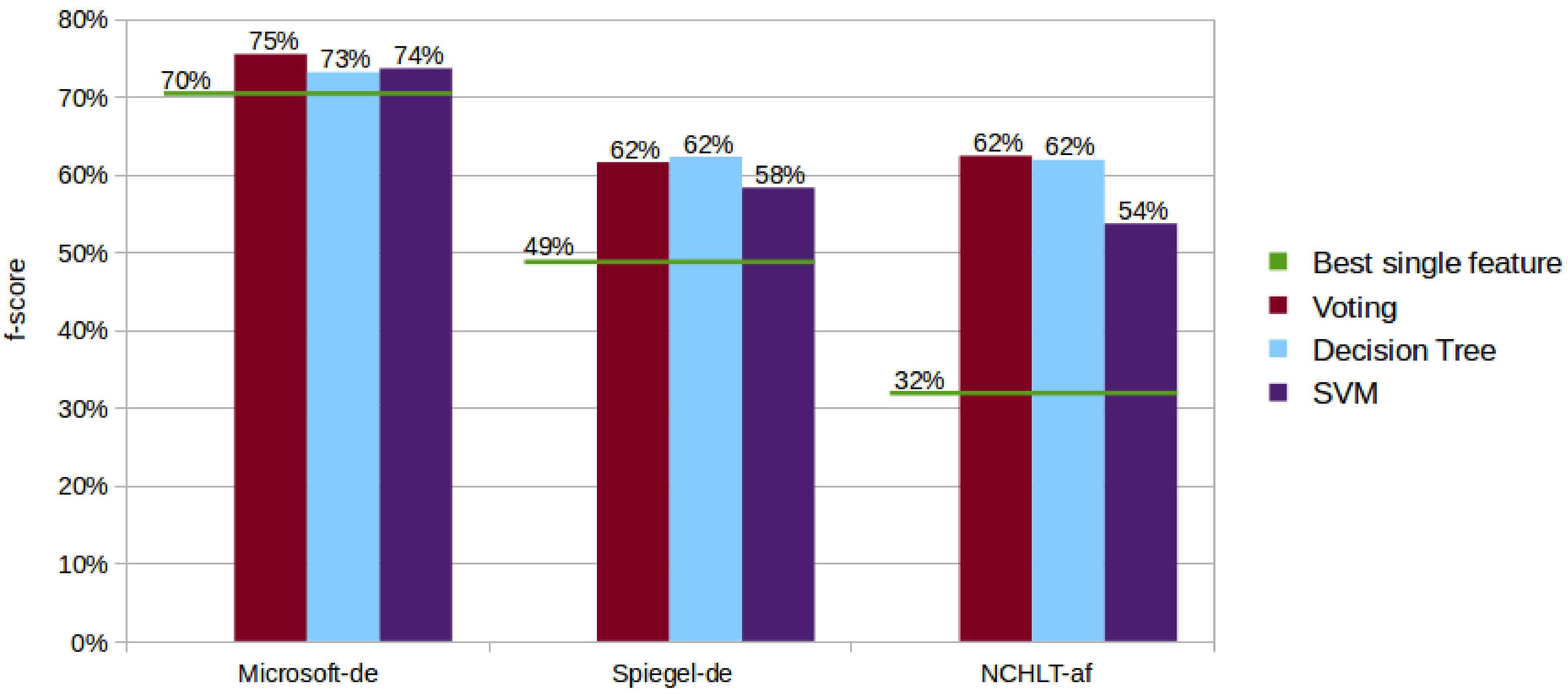


Result: Single Features



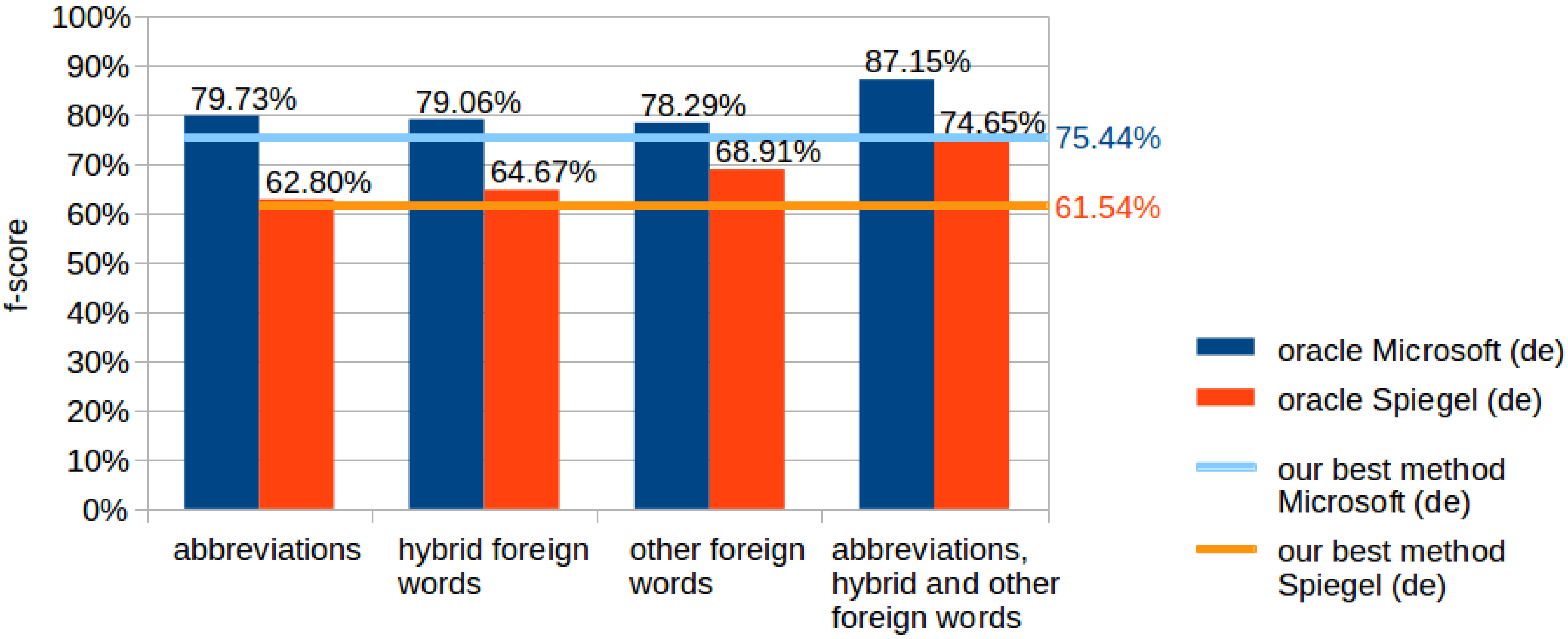
➔ On Spiegel-de test set: Higher ratio of words classified as English are wrong

Result: Combination



Challenges

- Performance after filtering difficult words (oracle)

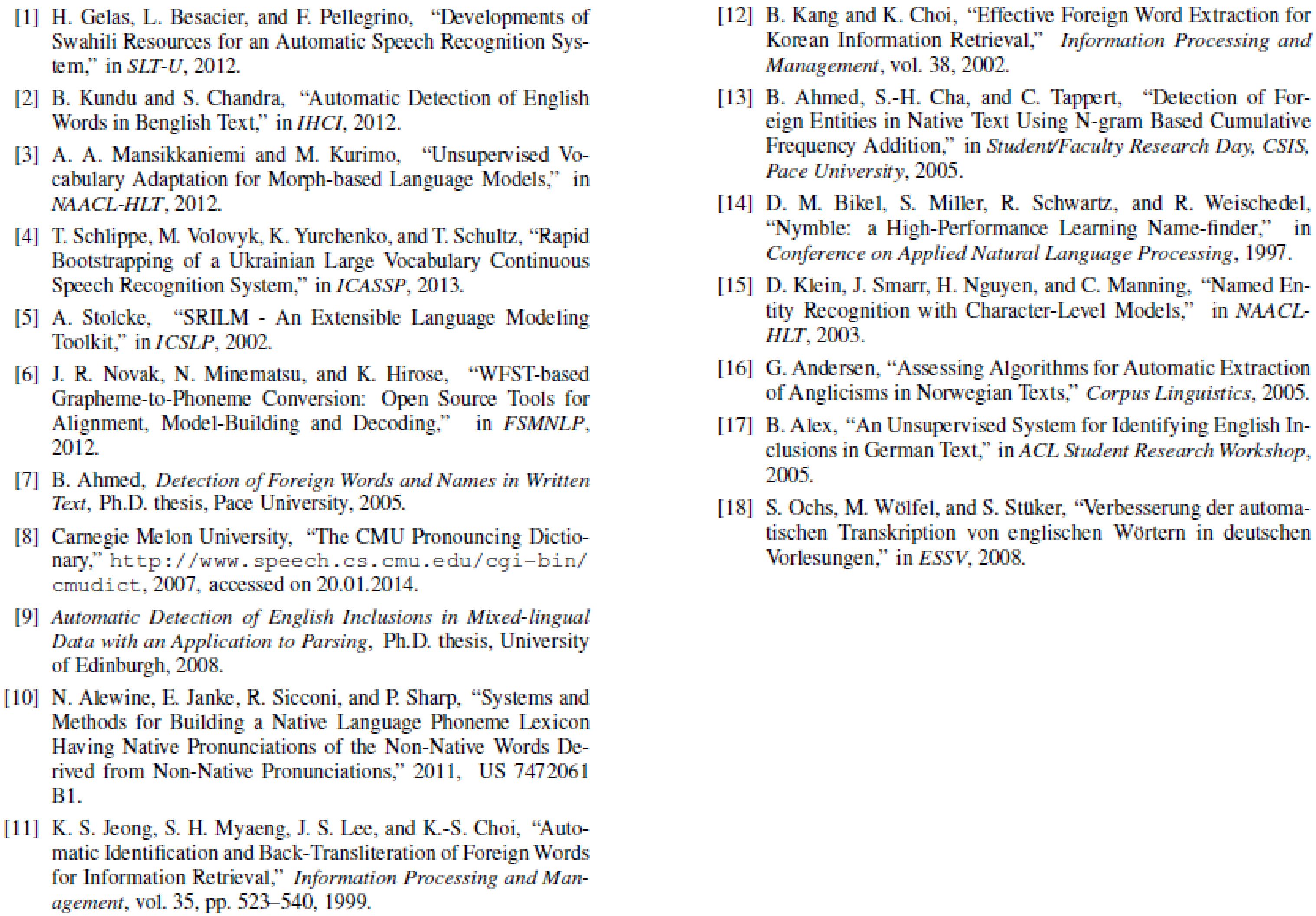


Conclusion and Future Work

- Features based on available sources
- New approaches:
  - G2P confidence
  - Wiktionary
- Further features:
  - Part-of-speech (POS)
  - Context, trigger words
  - Capitalization
  - Translate and compare

благодари́м за внима́ние!

References



[1] H. Gelas, L. Besacier, and F. Pellegrino, "Developments of Swahili Resources for an Automatic Speech Recognition System," in *SLT-U*, 2012.

[2] B. Kundu and S. Chandra, "Automatic Detection of English Words in Benglish Text," in *IHCI*, 2012.

[3] A. A. Mansikkaniemi and M. Kurimo, "Unsupervised Vocabulary Adaptation for Morph-based Language Models," in *NAACL-HLT*, 2012.

[4] T. Schlippe, M. Volovyk, K. Yurchenko, and T. Schultz, "Rapid Bootstrapping of a Ukrainian Large Vocabulary Continuous Speech Recognition System," in *ICASSP*, 2013.

[5] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *ICSLP*, 2002.

[6] J. R. Novak, N. Minematsu, and K. Hirose, "WFST-based Grapheme-to-Phoneme Conversion: Open Source Tools for Alignment, Model-Building and Decoding," in *FSMNLP*, 2012.

[7] B. Ahmed, *Detection of Foreign Words and Names in Written Text*, Ph.D. thesis, Pace University, 2005.

[8] Carnegie Melon University, "The CMU Pronouncing Dictionary," `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, 2007, accessed on 20.01.2014.

[9] *Automatic Detection of English Inclusions in Mixed-lingual Data with an Application to Parsing*, Ph.D. thesis, University of Edinburgh, 2008.

[10] N. Alewine, E. Janke, R. Sicconi, and P. Sharp, "Systems and Methods for Building a Native Language Phoneme Lexicon Having Native Pronunciations of the Non-Native Words Derived from Non-Native Pronunciations," 2011, US 7472061 B1.

[11] K. S. Jeong, S. H. Myaeng, J. S. Lee, and K.-S. Choi, "Automatic Identification and Back-Transliteration of Foreign Words for Information Retrieval," *Information Processing and Management*, vol. 35, pp. 523–540, 1999.

[12] B. Kang and K. Choi, "Effective Foreign Word Extraction for Korean Information Retrieval," *Information Processing and Management*, vol. 38, 2002.

[13] B. Ahmed, S.-H. Cha, and C. Tappert, "Detection of Foreign Entities in Native Text Using N-gram Based Cumulative Frequency Addition," in *Student/Faculty Research Day, CSIS, Pace University*, 2005.

[14] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a High-Performance Learning Name-finder," in *Conference on Applied Natural Language Processing*, 1997.

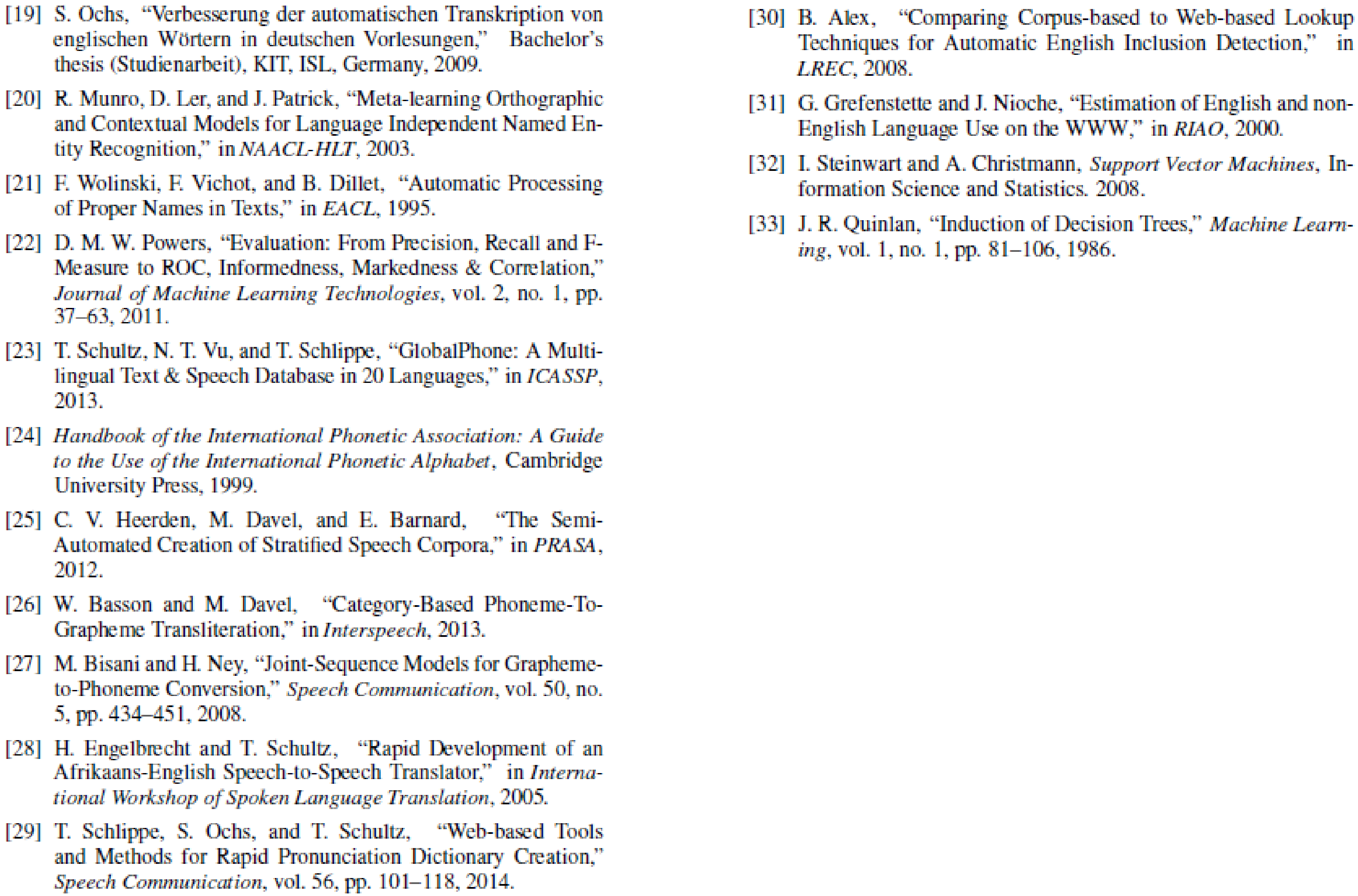
[15] D. Klein, J. Smarr, H. Nguyen, and C. Manning, "Named Entity Recognition with Character-Level Models," in *NAACL-HLT*, 2003.

[16] G. Andersen, "Assessing Algorithms for Automatic Extraction of Anglicisms in Norwegian Texts," *Corpus Linguistics*, 2005.

[17] B. Alex, "An Unsupervised System for Identifying English Inclusions in German Text," in *ACL Student Research Workshop*, 2005.

[18] S. Ochs, M. Wölfel, and S. Stüker, "Verbesserung der automatischen Transkription von englischen Wörtern in deutschen Vorlesungen," in *ESSV*, 2008.

References



[19] S. Ochs, "Verbesserung der automatischen Transkription von englischen Wörtern in deutschen Vorlesungen," Bachelor's thesis (Studienarbeit), KIT, ISL, Germany, 2009.

[20] R. Munro, D. Ler, and J. Patrick, "Meta-learning Orthographic and Contextual Models for Language Independent Named Entity Recognition," in *NAACL-HLT*, 2003.

[21] F. Wolinski, F. Vichot, and B. Dillet, "Automatic Processing of Proper Names in Texts," in *EACL*, 1995.

[22] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[23] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text & Speech Database in 20 Languages," in *ICASSP*, 2013.

[24] *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[25] C. V. Heerden, M. Davel, and E. Barnard, "The Semi-Automated Creation of Stratified Speech Corpora," in *PRASA*, 2012.

[26] W. Basson and M. Davel, "Category-Based Phoneme-To-Grapheme Transliteration," in *Interspeech*, 2013.

[27] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[28] H. Engelbrecht and T. Schultz, "Rapid Development of an Afrikaans-English Speech-to-Speech Translator," in *International Workshop of Spoken Language Translation*, 2005.

[29] T. Schlippe, S. Ochs, and T. Schultz, "Web-based Tools and Methods for Rapid Pronunciation Dictionary Creation," *Speech Communication*, vol. 56, pp. 101–118, 2014.

[30] B. Alex, "Comparing Corpus-based to Web-based Lookup Techniques for Automatic English Inclusion Detection," in *LREC*, 2008.

[31] G. Grefenstette and J. Nioche, "Estimation of English and non-English Language Use on the WWW," in *RIAO*, 2000.

[32] I. Steinwart and A. Christmann, *Support Vector Machines*, Information Science and Statistics. 2008.

[33] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.