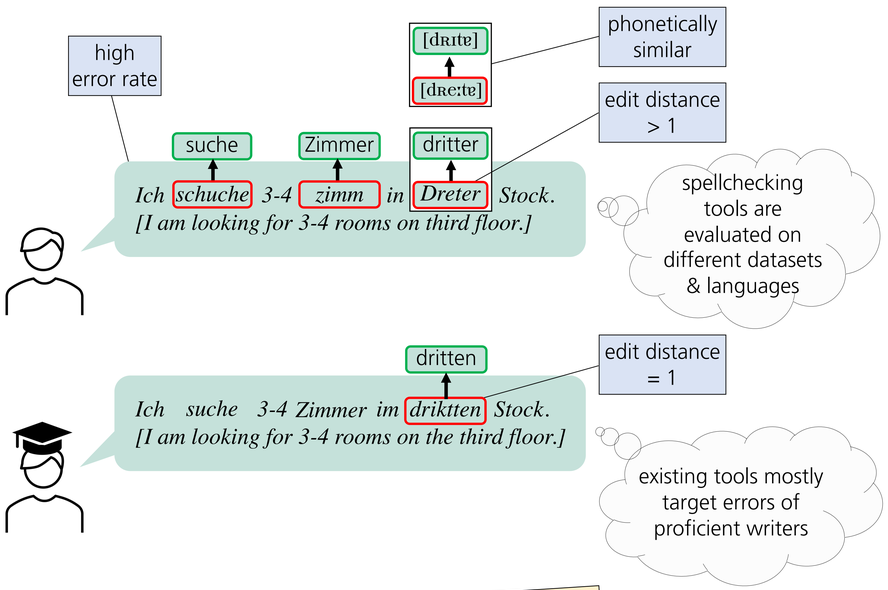
LeSpell – A Multi-Lingual Benchmark Corpus of Spelling Errors to Develop Spellchecking Methods for Learner Language

Marie Bexte, Ronja Laarmann-Quante, Andrea Horbach, Torsten Zesch
Research Cluster "Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education", FernUniversität in Hagen, Germany

Motivation: Misspellings in Learner Texts

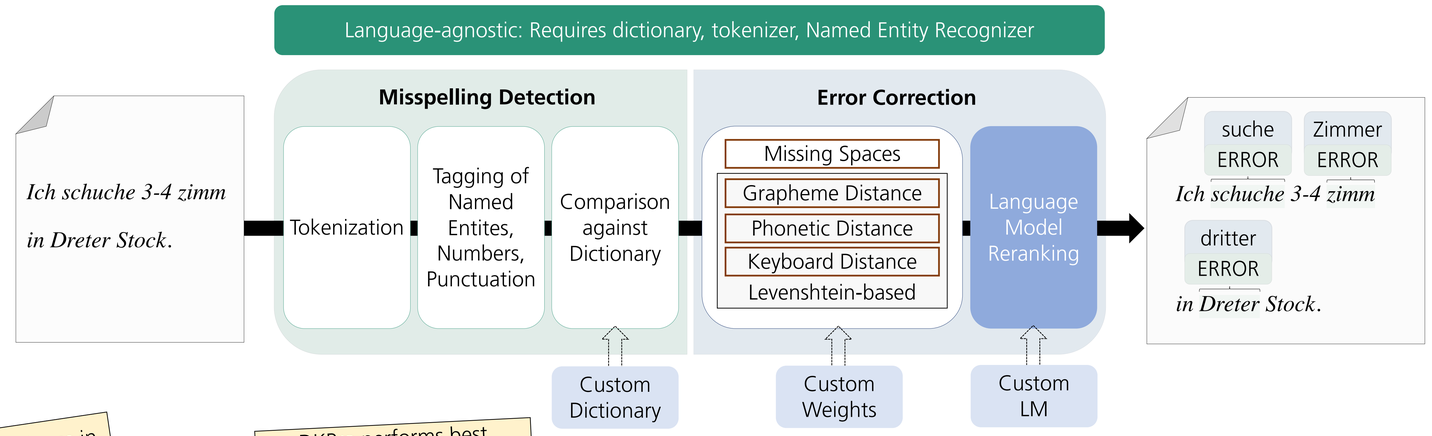


The LeSpell Data Set

Uniform XML format for contextualized erros to allow for easy benchmarking experiments

		Modality	# Errors	Ø Levensht.	Error Rate	
TOEFL-Spell (Flor et al., 2019)	L2	⌨	6,251	1.25	2.19%	🇺🇸
MERLIN-CZ (Wisniewski et al., 2018)	L2	✎	4,807	1.35	9.95%	🇨🇿
CItA (Barbagli et al., 2016)	L1	✎	1,431	1.14	0.39%	🇮🇹
MERLIN-IT (Wisniewski et al., 2018)	L2	✎	2,137	1.18	3.22%	
Litkey (Laarmann-Quante et al., 2019)	L1	✎	38,698	1.40	20.20%	🇩🇪
MERLIN-DE (Wisniewski et al., 2018)	L2	✎	5,366	1.34	6.94%	
SkaLa (Scholten-Akoun et al., 2014)	L1	⌨	423	1.25	3.40%	

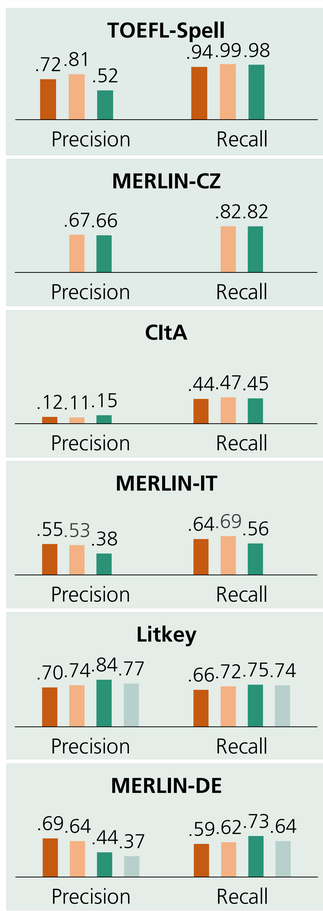
DKPro Spelling



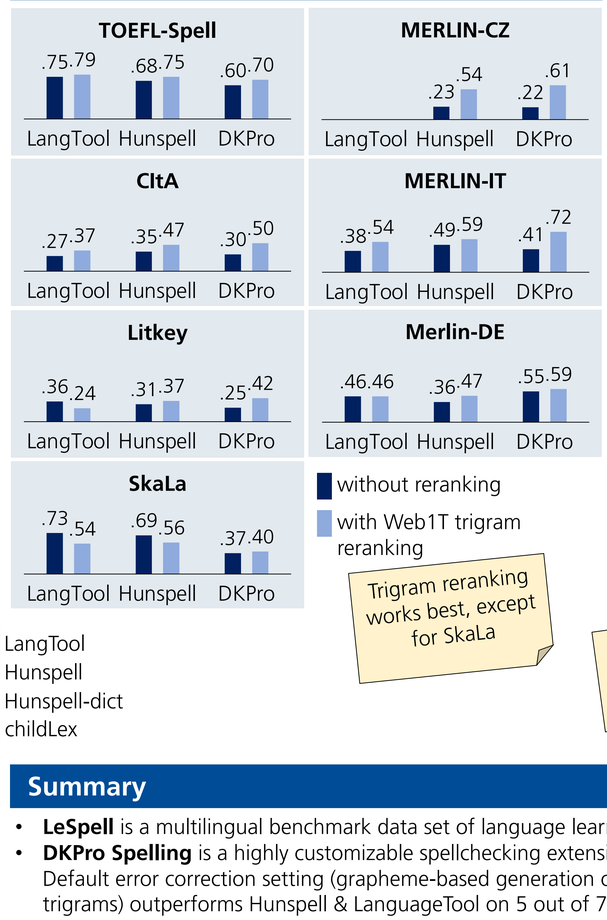
Large variance in performance

DKPro performs best on all subcorpora except TOEFL-Spell & SkaLa

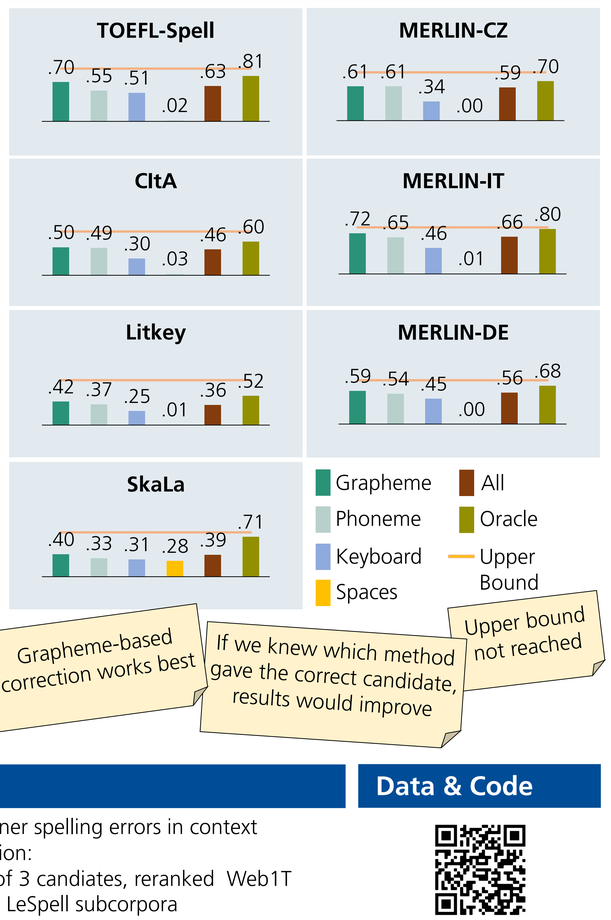
Misspelling Detection



Correction Recall@1: Comparison of Tools



Correction Recall@1: Comparison of Methods



Trigram reranking works best, except for SkaLa

Grapheme-based correction works best

If we knew which method gave the correct candidate, results would improve

Upper bound not reached

Summary

- **LeSpell** is a multilingual benchmark data set of language learner spelling errors in context
- **DKPro Spelling** is a highly customizable spellchecking extension:
  Default error correction setting (grapheme-based generation of 3 candiates, reranked Web1T trigrams) outperforms Hunspell & LanguageTool on 5 out of 7 LeSpell subcorpora

Data & Code

This work was conducted in the framework of the Research Cluster D²L² "Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education" of the FernUniversität in Hagen, Germany (https://e.feu.de/english-d2l2)