Candidate Evaluation Strategies for Improved Difficulty Prediction of Language Proficiency Tests

BEA Workshop 2015
Lisa Beinborn, Torsten Zesch, Iryna Gurevych
UKP Lab
Department of Computer Science
Technische Universität Darmstadt



Follow-up work

TECHNISCHE UNIVERSITÄT DARMSTADT

TACL paper 2014:

“Predicting the Difficulty of Language Proficiency Tests”

What is difficult for language learners? (And why?)

How can we predict difficulty?

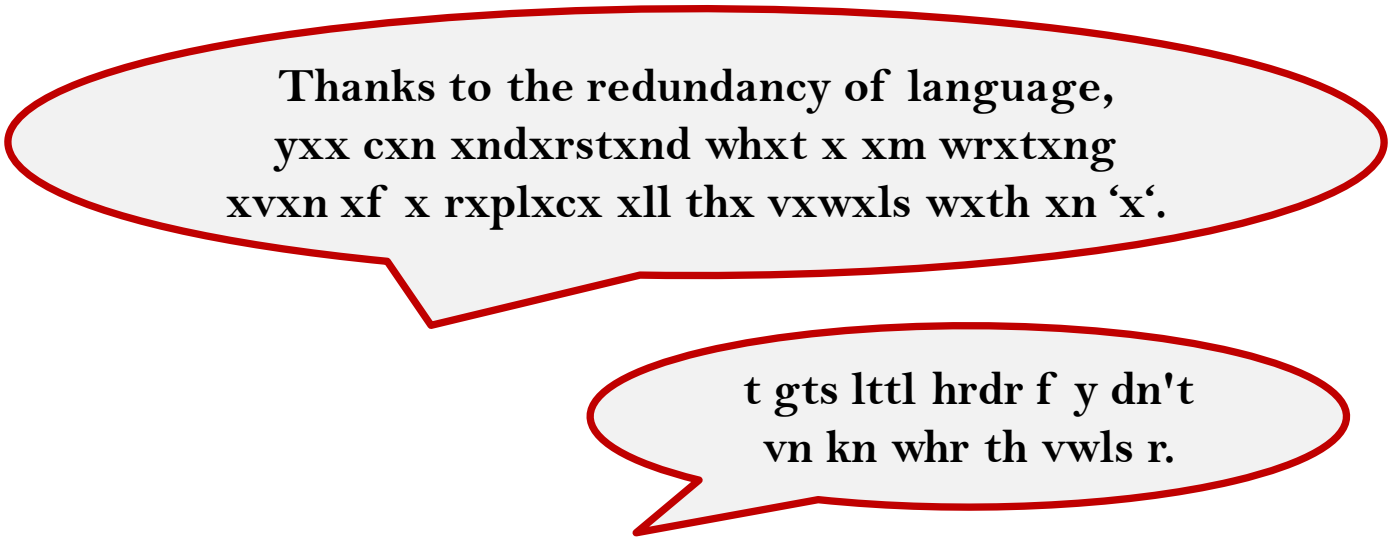
Data: Language Proficiency Tests

Task: Predict difficulty of the items

Reduced Redundancy Principle

TECHNISCHE UNIVERSITÄT DARMSTADT

- Spolsky (1969): “Natural language is redundant”
- The ability to deal with reduced redundancy distinguishes beginners from advanced language learners



Steven Pinker, The Language Instinct: How the Mind Creates Language (William Morrow, 1994)

UKP

C-Test [Klein-Braley and Raatz, 1982]

TECHNISCHE UNIVERSITÄT DARMSTADT

- Beginning and end of text provide context
- Every second word is a gap
- Smaller “half” of the word is provided

The roots of humanity can be traced back to millions of years ago. T____ primary evid_____ comes fr_____ fossils - skulls, skel_____and bo_____ fragments. Scien_____ have ma_____ tools th_____ allow th_____ to ext_____ subtle infor_____ from anc_____ bones a_____ their enviro_____ settings. Mod_____forensic wo_____ in t_____ field a_____ in labora_____ can n_____ provide a rich understanding of how our ancestors lived.

UKP

C-Test [Klein-Braley and Raatz, 1982]

TECHNISCHE UNIVERSITÄT DARMSTADT

- Beginning and end of text provide context
- Every second word is a gap
- Smaller "half" of the word is provided

The roots of humanity can be traced back to millions of years ago. T____ primary evid_____ comes fr_____ fossils - skulls, skel_____and bo_____ fragments. Scien_____ have ma_____ tools th_____ allow th_____ to ext_____ subtle infor_____ from anc_____ bones a_____ their enviro_____ settings. Mod_____forensic wo_____ in t_____ field a_____ in labora_____ can n_____ provide a rich understanding of how our ancestors lived.

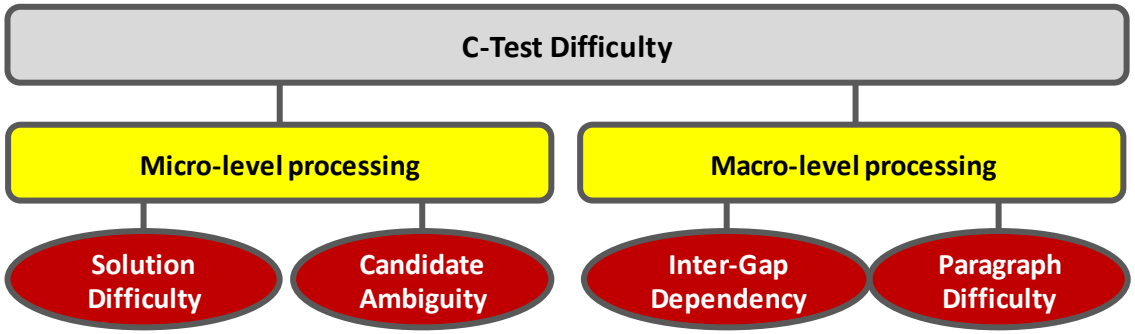
Data: TU Darmstadt Placement test at language centre

UKP

Yesterday:
Difficulty Prediction of English C-test Gaps

TECHNISCHE UNIVERSITÄT DARMSTADT

The roots of humanity can be traced back to millions of years ago. T● primary evid● comes fr● fossils - skulls, skel● and bo● fragments. Scien● have ma● tools th● allow th● to ext● subtle infor● from anc● bones a● their enviro● settings. Mod● forensic wo● in t● field a● in labora● can n● provide a rich understanding of how our ancestors lived.



UKP

Outlook of TACL paper

TECHNISCHE UNIVERSITÄT DARMSTADT

1) Adapt to other languages and to other test variants
2) Improve the dimensions candidate ambiguity and inter-gap dependency

TODAY: Part 1

TECHNISCHE UNIVERSITÄT DARMSTADT

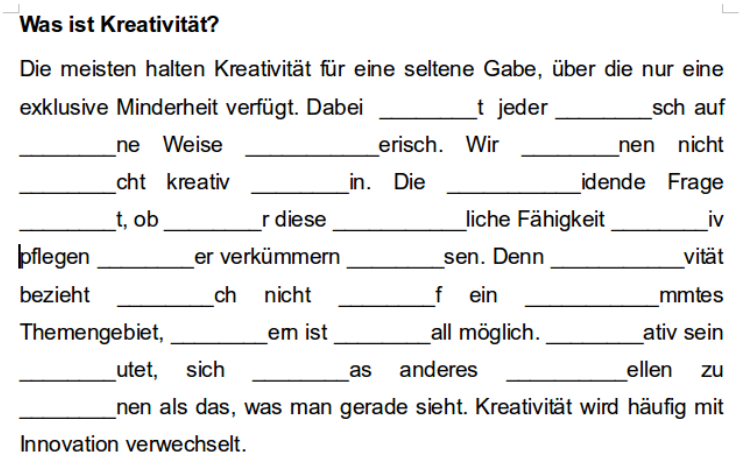
1) Adapt to **German and French** and to **two test variants**
2) Improve the dimensions candidate ambiguity and inter-gap dependency

TODAY, Part 2

TECHNISCHE UNIVERSITÄT DARMSTADT

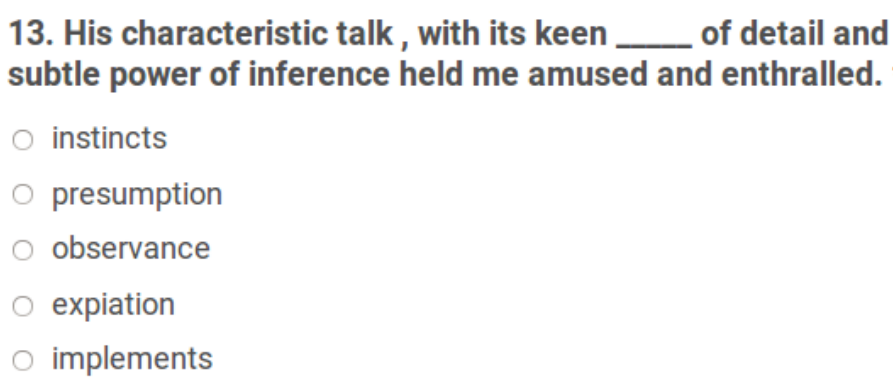
1) Adapt to German and French and to two **test variants**

2) Improve the dimensions **candidate ambiguity** and inter-gap dependency



Was ist Kreativität?

Die meisten halten Kreativität für eine seltene Gabe, über die nur eine exklusive Minderheit verfügt. Dabei _____t jeder _____sch auf _____ne Weise _____erisch. Wir _____nen nicht _____cht kreativ _____in. Die _____idende Frage _____t, ob _____r diese _____liche Fähigkeit _____iv pflegen _____er verkümmern _____sen. Denn _____vität bezieht _____ch nicht _____f ein _____mmtes Themengebiet, _____em ist _____all möglich. _____ativ sein _____utet, sich _____as anderes _____ellen zu _____nen als das, was man gerade sieht. Kreativität wird häufig mit Innovation verwechselt.



13. His characteristic talk , with its keen _____ of detail and subtle power of inference held me amused and enthralled.

- instincts
- presumption
- observance
- expiation
- implements

Different Languages

TECHNISCHE UNIVERSITÄT DARMSTADT

Is it a language-independent approach?

Can it be adapted to other languages?

Markus Koljonen: iki.fi/markus.koljonen

French C-Test

TECHNISCHE UNIVERSITÄT DARMSTADT

Le Noël des familles
Noël semble immuable, et pourtant il change ! Au co____ du te____ on s'aper____, discrètement ma____ sûrement, q____ la fê____ de l____ Nativité pr____ de nouv____ allures e____ alliant l____ traditions e____ la mode____. Traditionnellement, l____ sapin d____ Noël e____ surtout déc____ de bou____, de guirl____ lumineuses e____ de boules multicolores. Toutefois, une certaine influence germanique se fait de plus en plus sentir, avec la présence de figurines en bois ou de dessins aux fenêtres.

- articles
- accents
- richer morphology

UKP

French C-Test

TECHNISCHE UNIVERSITÄT DARMSTADT

Le Noël des familles
Noël semble immuable, et pourtant il change ! Au co____ du te____ on s'aper____, discrètement ma____ sûrement, q____ la fê____ de l____ Nativité pr____ de nouv____ allures e____ alliant l____ traditions e____ la mode____. Traditionnellement, l____ sapin d____ Noël e____ surtout déc____ de bou____, de guirl____ lumineuses e____ de boules multicolores. Toutefois, une certaine influence germanique se fait de plus en plus sentir, avec la présence de figurines en bois ou de dessins aux fenêtres.

Data: TU Darmstadt Placement test at language centre

German C-Test

TECHNISCHE UNIVERSITÄT DARMSTADT

Auf einer Weltkarte kann man sehen, dass Asien und Nordamerika im Norden nur durch einen schmalen Meeresstreifen voneinander getrennt sind: durch die Beringstraße. Während d_____ Eiszeit herrs_____ auf d_____ ganzen Er_____ niedrigere Temper_____ , und d_____ Beringstraße w_____ zugefroren. Üb_____ das E_____ gelangten Volksg_____ aus As_____ auf d_____ amerikanischen Kont_____ . Manche bli_____ in Norda_____ und bild_____ mehr a_____ tausend versch_____ Stämme - jew_____ mit ei_____ eigenen Sprache, die anderen zogen weiter bis nach Südamerika.

- articles, cases
- Umlaute, case sensitivity
- old/new orthography
- compounds

UKP

German C-Test

TECHNISCHE UNIVERSITÄT DARMSTADT

Auf einer Weltkarte kann man sehen, dass Asien und Nordamerika im Norden nur durch einen schmalen Meeresstreifen voneinander getrennt sind: durch die Beringstraße. Während d_____ Eiszeit herrs_____ auf d_____ ganzen Er_____ niedrigere Temper_____ , und d_____ Beringstraße w_____ zugefroren. Üb_____ das E_____ gelangten Volksg_____ aus As_____ auf d_____ amerikanischen Kont_____ . Manche bli_____ in Norda_____ und bild_____ mehr a_____ tausend versch_____ Stämme - jew_____ mit ei_____ eigenen Sprache, die anderen zogen weiter bis nach Südamerika.

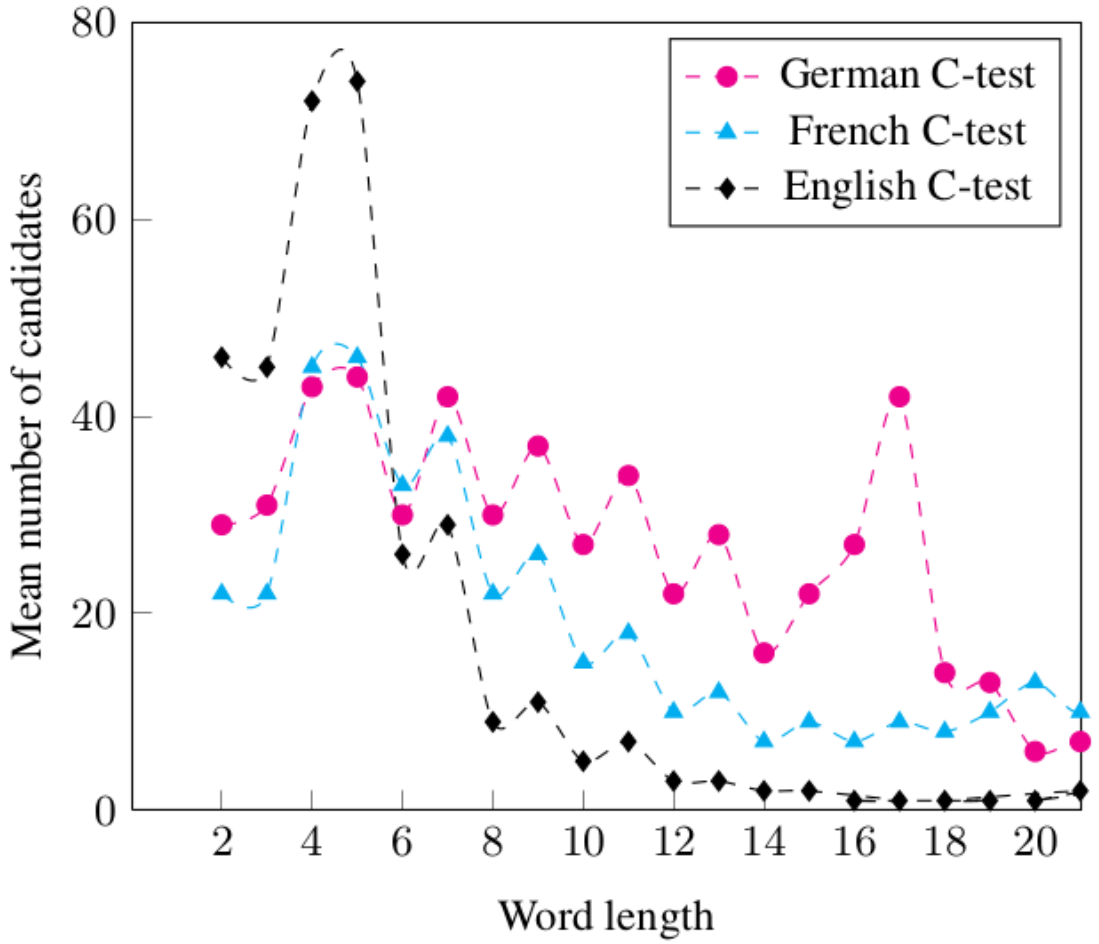
Data: TestDaF Institute Certificate of German proficiency for university admission

Different Languages...

TECHNISCHE UNIVERSITÄT DARMSTADT

Language	Words	Mean word length
English	99,171	8.5 ± 2.6
French	139,719	9.6 ± 2.6
German	332,263	12.0 ± 3.5

Different Candidate Space



Different Test Types

TECHNISCHE UNIVERSITÄT DARMSTADT

The reduced redundancy principle comprises more than just the C-test.

Prefix Deletion Test

TECHNISCHE UNIVERSITÄT DARMSTADT

Was ist Kreativität?
Die meisten halten Kreativität für eine seltene Gabe, über die nur eine exklusive Minderheit verfügt. Dabei _____t jeder _____sch auf _____ne Weise _____erisch. Wir _____nen nicht _____cht kreativ _____in. Die _____idende Frage _____t, ob _____r diese _____liche Fähigkeit _____iv pflegen _____er verkümmern _____sen. Denn _____vität bezieht _____ch nicht _____f ein _____mmtes Themengebiet, _____ern ist _____all möglich. _____ativ sein _____utet, sich _____as anderes _____ellen zu _____nen als das, was man gerade sieht. Kreativität wird häufig mit Innovation verwechselt.

- **prefix vs postfix**
- **multiple solutions**

UKP

Prefix Deletion Test

TECHNISCHE UNIVERSITÄT DARMSTADT

Was ist Kreativität?
Die meisten halten Kreativität für eine seltene Gabe, über die nur eine exklusive Minderheit verfügt. Dabei _____t jeder _____sch auf _____ne Weise _____erisch. Wir _____nen nicht _____cht kreativ _____in. Die _____idende Frage _____t, ob _____r diese _____liche Fähigkeit _____iv pflegen _____er verkümmern _____sen. Denn _____vität bezieht _____ch nicht _____f ein _____mmtes Themengebiet, _____ern ist _____all möglich. _____ativ sein _____utet, sich _____as anderes _____ellen zu _____nen als das, was man gerade sieht. Kreativität wird häufig mit Innovation verwechselt.

Data: University of Duisburg-Essen German proficiency test for prospective teachers

Cloze tests

TECHNISCHE UNIVERSITÄT DARMSTADT

13. His characteristic talk , with its keen _____ of detail and subtle power of inference held me amused and enthralled.

- instincts
- presumption
- observance
- expiation
- implements

- **Closed format**
- **Distractors**

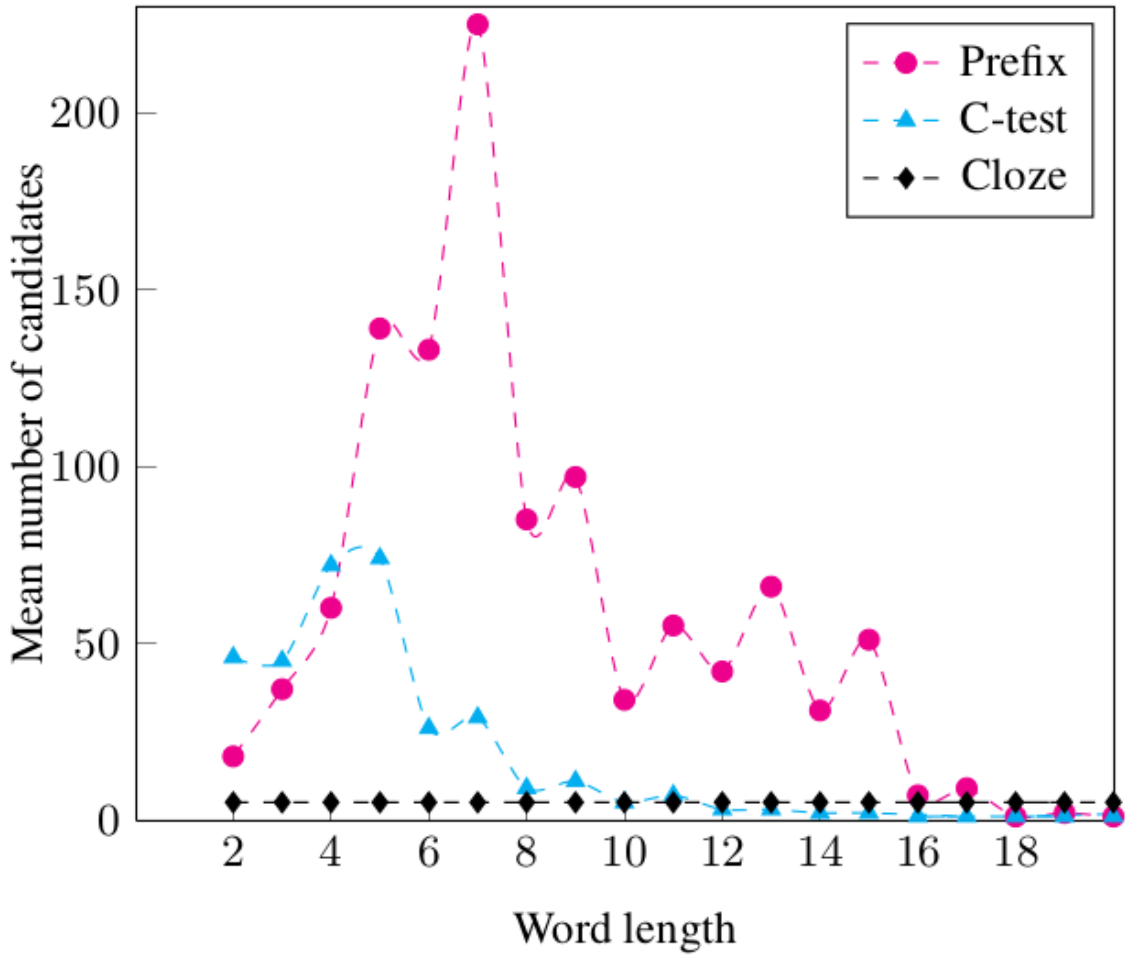
Cloze tests

13. His characteristic talk , with its keen _____ of detail and subtle power of inference held me amused and enthralled.

- instincts
- presumption
- observance
- expiation
- implements

Questions:
Microsoft Research
Zweig & Burges (2012)
Error Rates: own study

Candidate Space



Data

TECHNISCHE UNIVERSITÄT DARMSTADT

Format	Test Type	Texts	Gaps	Participants	Av. Error Rate
Open	C-test en	39	775	210	.35 ± .25
	C-test fr	40	799	24	.52 ± .28
	C-test de	82	1,640	251	.55 ± .26
	Prefix de	14	348	225	.36 ± .23
Closed	Cloze en	100	100	22	.27 ± .22

Generalization

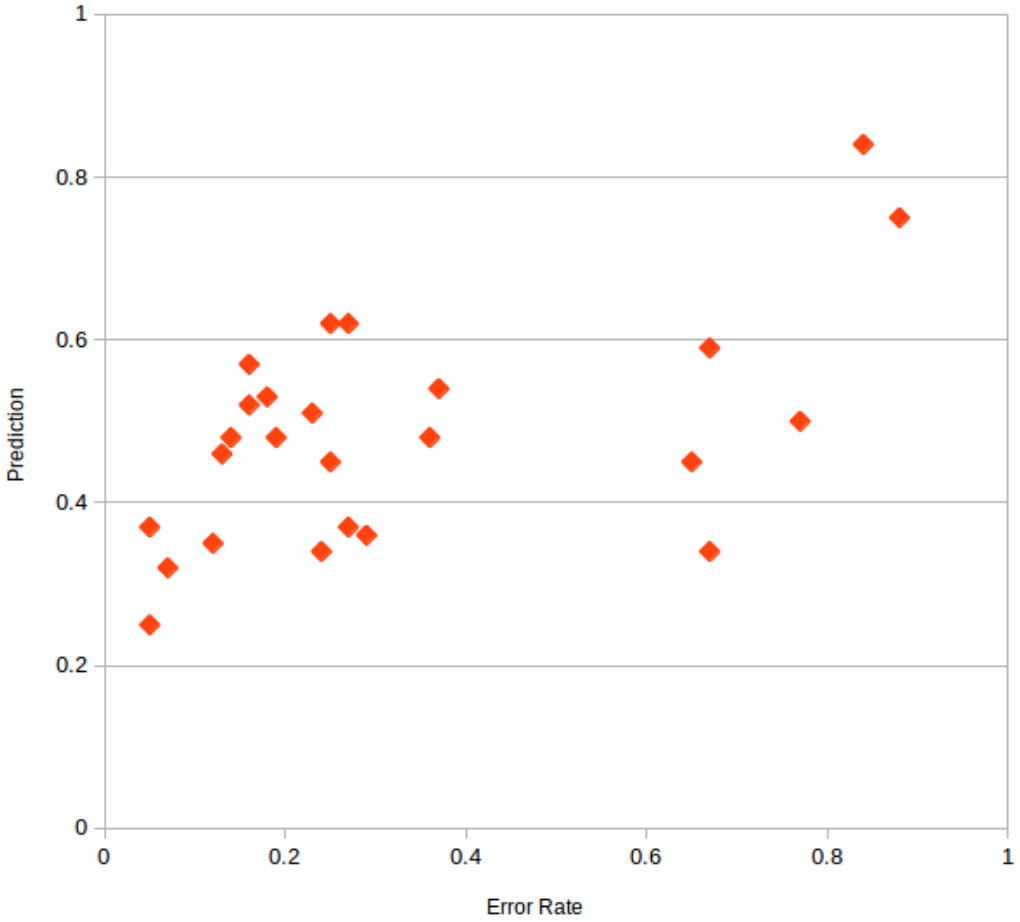
TECHNISCHE UNIVERSITÄT DARMSTADT

- Adapt format and types
- Adapt linguistic pre-processing (DKPro)
- Adapt resources
- Adapt features (Reduction from 87 to 70)

Prediction Task

TECHNISCHE UNIVERSITÄT DARMSTADT

- SMO regression, Leave-one-out cross-validation on texts



UKP

Prediction Results: Languages

TECHNISCHE UNIVERSITÄT DARMSTADT

Test Type	Pearson's r
C-test en	.47
C-test fr	.67
C-test de	.61

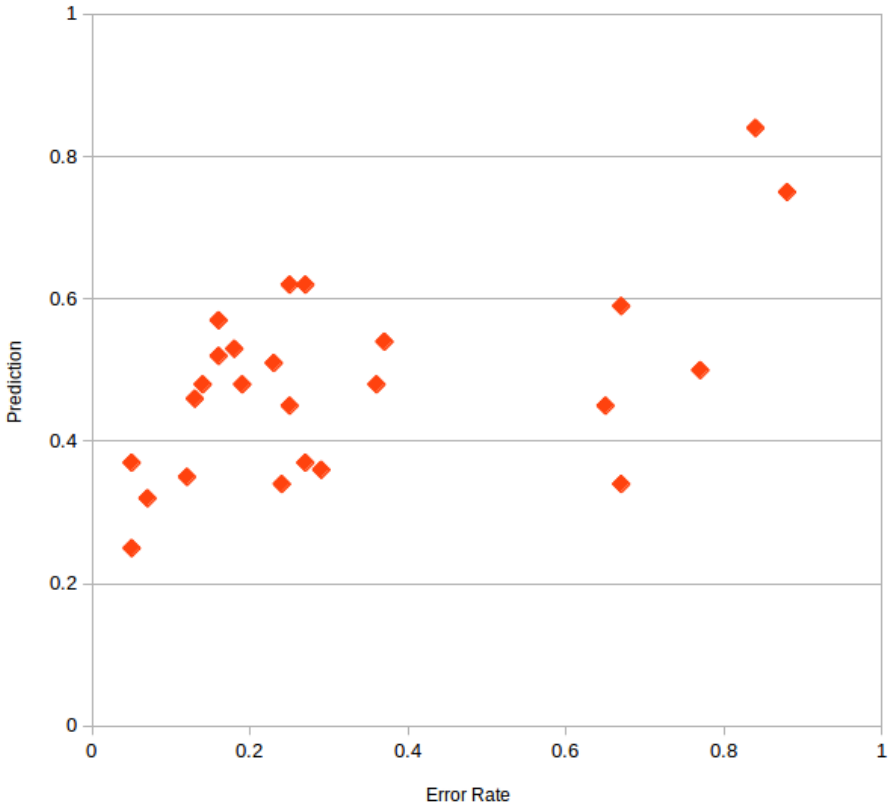
Prediction Results

TECHNISCHE UNIVERSITÄT DARMSTADT

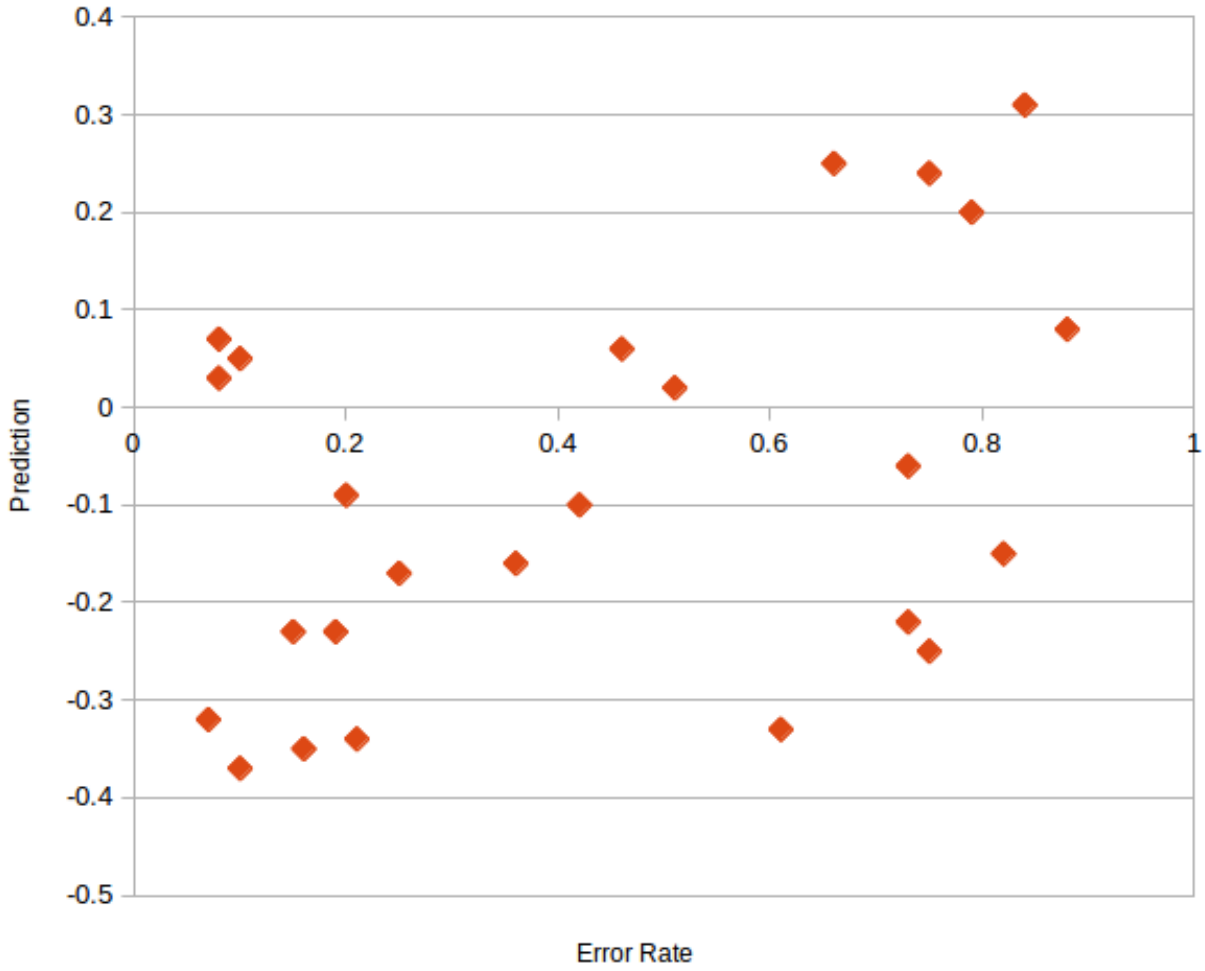
Test Type	Pearson's r
C-test en	.47
C-test fr	.67
C-test de	.61
Prefix de	.27
Cloze en	.20

Prediction Results

Test Type	Pearson's r
C-test en	.47
C-test fr	.67
C-test de	.61
Prefix de	.27
Cloze en	.20



Prefix Deletion: Outlier



Open vs closed format

- Many of the features developed for difficulty prediction were targeted at production problems (e.g. spelling).
- Closed cloze tests only require recognition skills.

Test Type	Pearson's r
C-test en	.47
C-test fr	.67
C-test de	.61
Prefix de	.27
Cloze en	.20

A closer look at cloze

TECHNISCHE UNIVERSITÄT DARMSTADT

49. The stage lost a fine _____ , even as science lost an acute reasoner , when he became a specialist in crime.

- linguist
- hunter
- actor
- estate
- horseman

Error Rate: 0.0

21. When his body had been carried from the cellar we found ourselves still confronted with a problem which was almost as _____ as that with which we had started.

- tall
- loud
- invisible
- quick
- formidable

Error Rate: 0.6

Inspiration from automated solving

TECHNISCHE UNIVERSITÄT DARMSTADT

How would we proceed for nlp-based solving of language tests?

Rank the candidates

1) According to language model probability
2) According to semantic relatedness between candidate and context

Differences

TECHNISCHE UNIVERSITÄT DARMSTADT

- Automatic solving:

  train on domain-specific data (Holmes novels)
- Difficulty prediction:
  train on general data to model learner knowledge

Language model: trained with Berkeley LM on Leipzig corpora

- 1 million sentences, news domain

Explicit Semantic Analysis Index: trained on Wikipedia

- 931,559 concepts

LM ranker

TECHNISCHE UNIVERSITÄT DARMSTADT

- Candidate fitness: log probability of sentence in language model

The stage lost a fine _____ , even as science lost an acute reasoner, when he became a specialist in crime .

actor	-358.83
estate	-361.22
hunter	-361.96
linguist	-362.71
horseman	-362.93

Rank of solution: **1**

LM ranker

TECHNISCHE UNIVERSITÄT DARMSTADT

- Candidate fitness: log probability of sentence in language model

When his body had been carried from the cellar we found ourselves still confronted with a problem which was almost as _____ as that with which we had started .

tall	-175.17
quick	-176.61
loud	-178.60
formidable	-179.52
invisible	-179.52

Rank of solution: **4**

ESA ranker

TECHNISCHE UNIVERSITÄT DARMSTADT

- Candidate fitness: sum over the cosine similarity between the candidate and every content word in the sentence (similar to Zweig et al. 2012)

The stage lost a fine _____ , even as science lost an acute reasoner, when he became a specialist in crime .

actor	1.06
estate	0.70
hunter	0.58
horseman	0.29
linguist	0.29

Rank of solution: **1**

ESA ranker

TECHNISCHE UNIVERSITÄT DARMSTADT

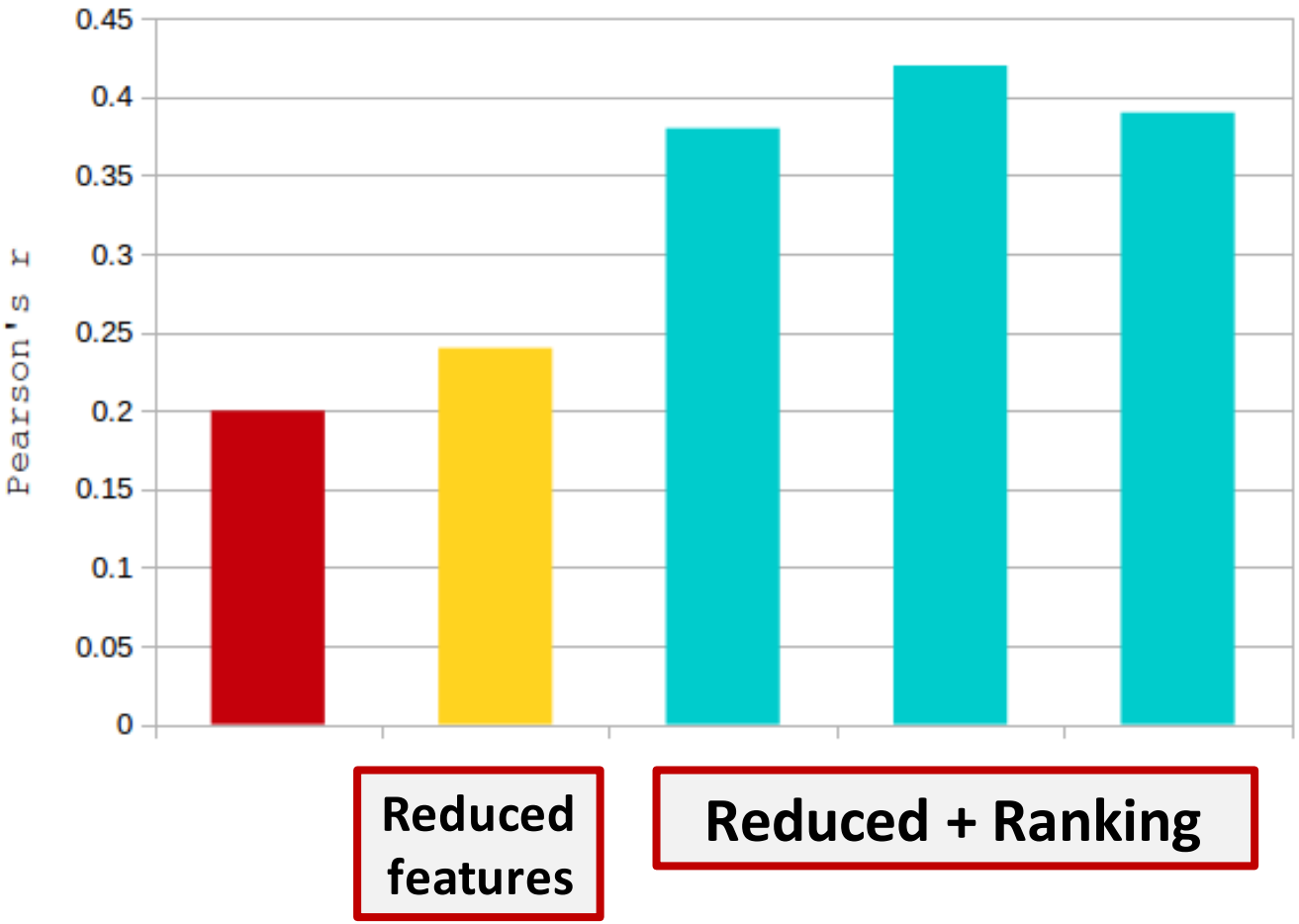
- Candidate fitness: sum over the cosine similarity between the candidate and every content word in the sentence (similar to Zweig et al. 2012)

When his body had been carried from the cellar we found ourselves still confronted with a problem which was almost as _____ as that with which we had started.

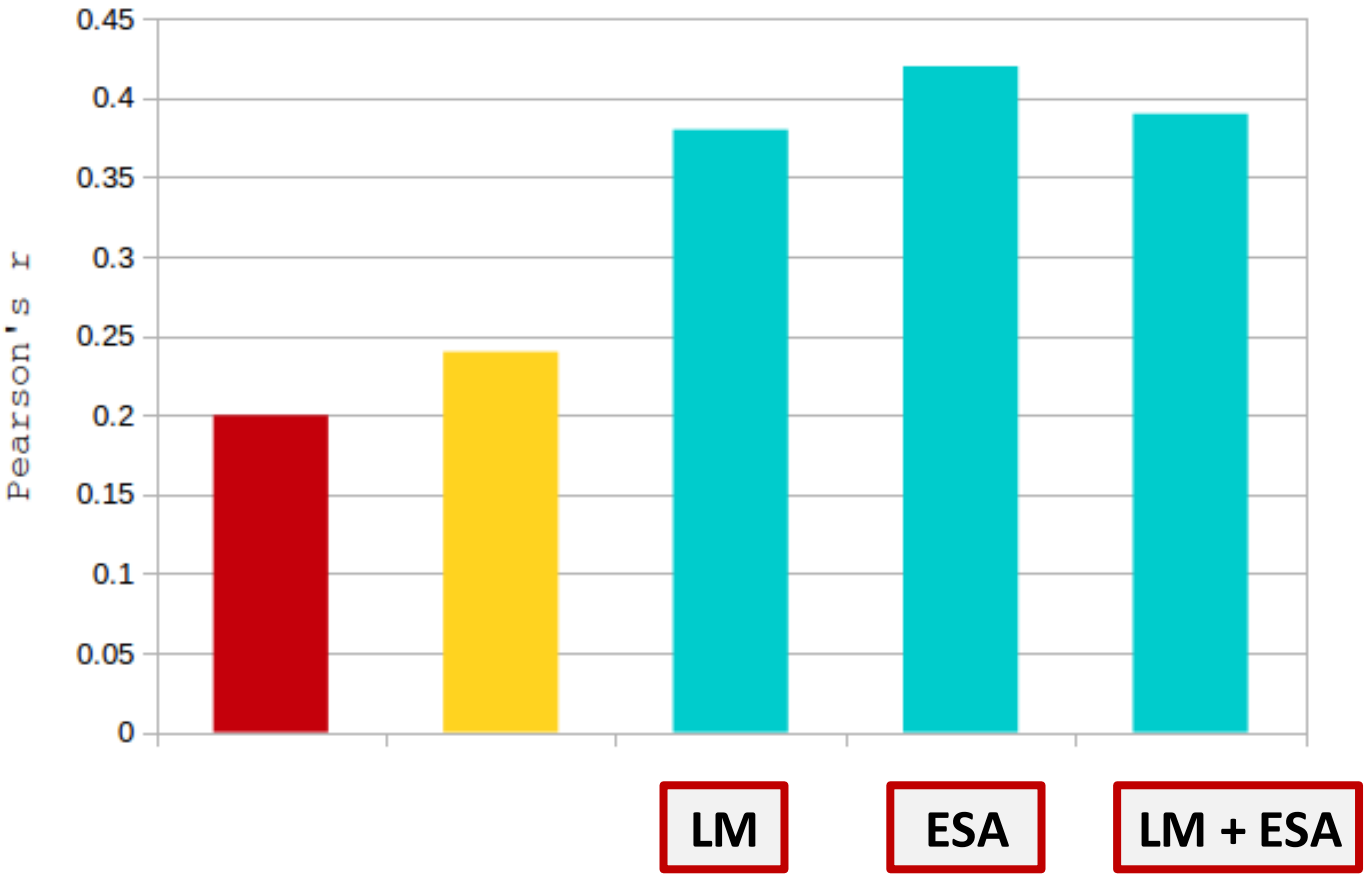
quick	1.03
tall	0.80
invisible	0.63
loud	0.61
formidable	0.56

Rank of solution: **5**

Improved Cloze Results



Improved Cloze Results



Summary and Outlook

TECHNISCHE UNIVERSITÄT DARMSTADT

- Succesfully adapted the prediction framework to new languages and test types.
- We  French.
- Candidate evaluation strategies from automatic solving can be applied to simulate learner difficulties.

Plans:

- Currently collecting more error rates for cloze tests.
- Use candidate evaluation strategies for distractor generation and difficulty manipulation.
- Strategies for inter-item dependencies.

Thank you very much for your _____.

- attention
- presence
- patience
- enthusiasm

Do you have any_____?

- questions
- comments
- praises
- doubts