Evaluating Morphosyntactic Wellformedness of Text

Adithya Pratapa,* Antonios Anastasopoulos,* Shruti Rijhwani,
Aditi Chaudhary, David R. Mortensen, Graham Neubig and Yulia Tsvetkov

EMNLP 2021

Carnegie Mellon University
Language
Technologies
Institute

Evaluating Language Generation

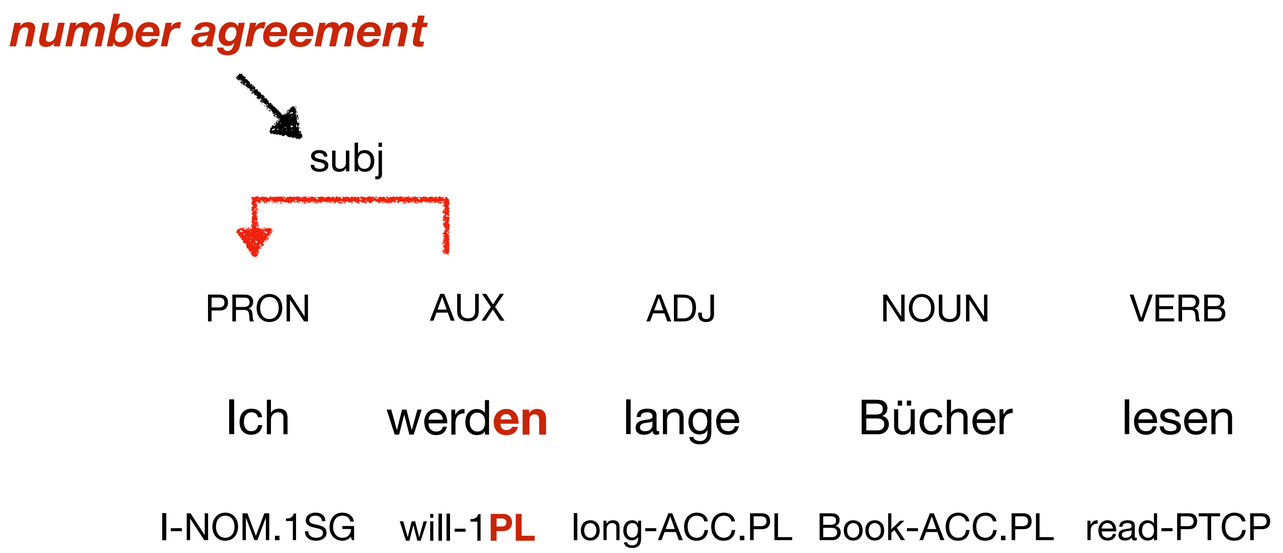
- Rapid progress in language generation in recent years
- A need for an interpretable evaluation of output text quality
- Metrics should be usable across many human languages

We propose a linguistically grounded evaluation of wellformedness in multilingual settings,

L'AMBRE

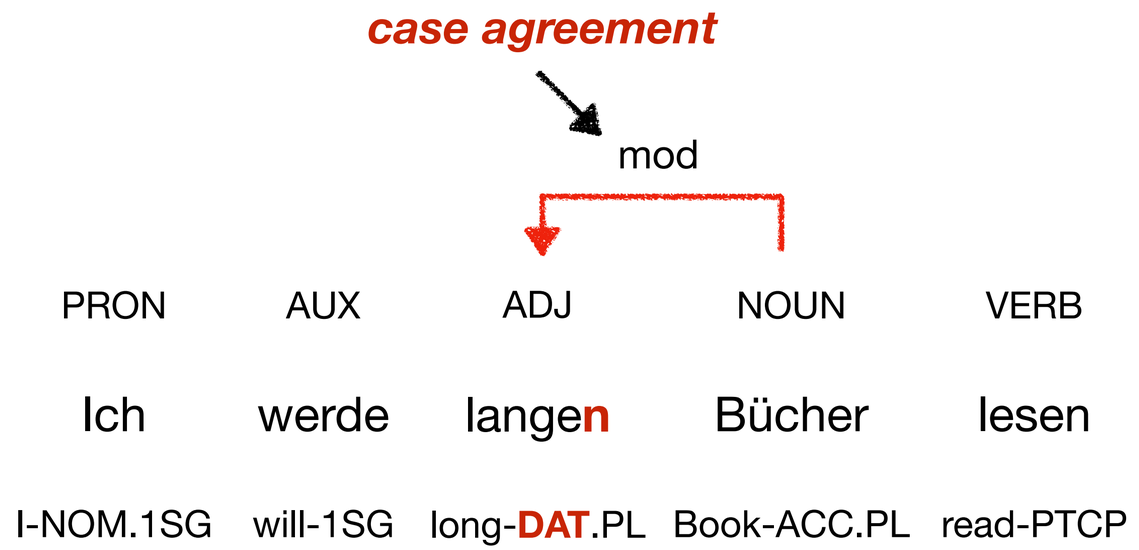
Grammatical Wellformedness

(*) Ich werd**en** lange Bücher lesen	Ich werde lange Bücher lesen
(*) Ich werde langen Bücher lesen	Ich werde lange Bücher lesen



Grammatical Wellformedness

(*) Ich werd**en** lange Bücher lesen	Ich werde lange Bücher lesen
(*) Ich werde lange**n** Bücher lesen	Ich werde lange Bücher lesen



Wellformedness ⊂ Fluency

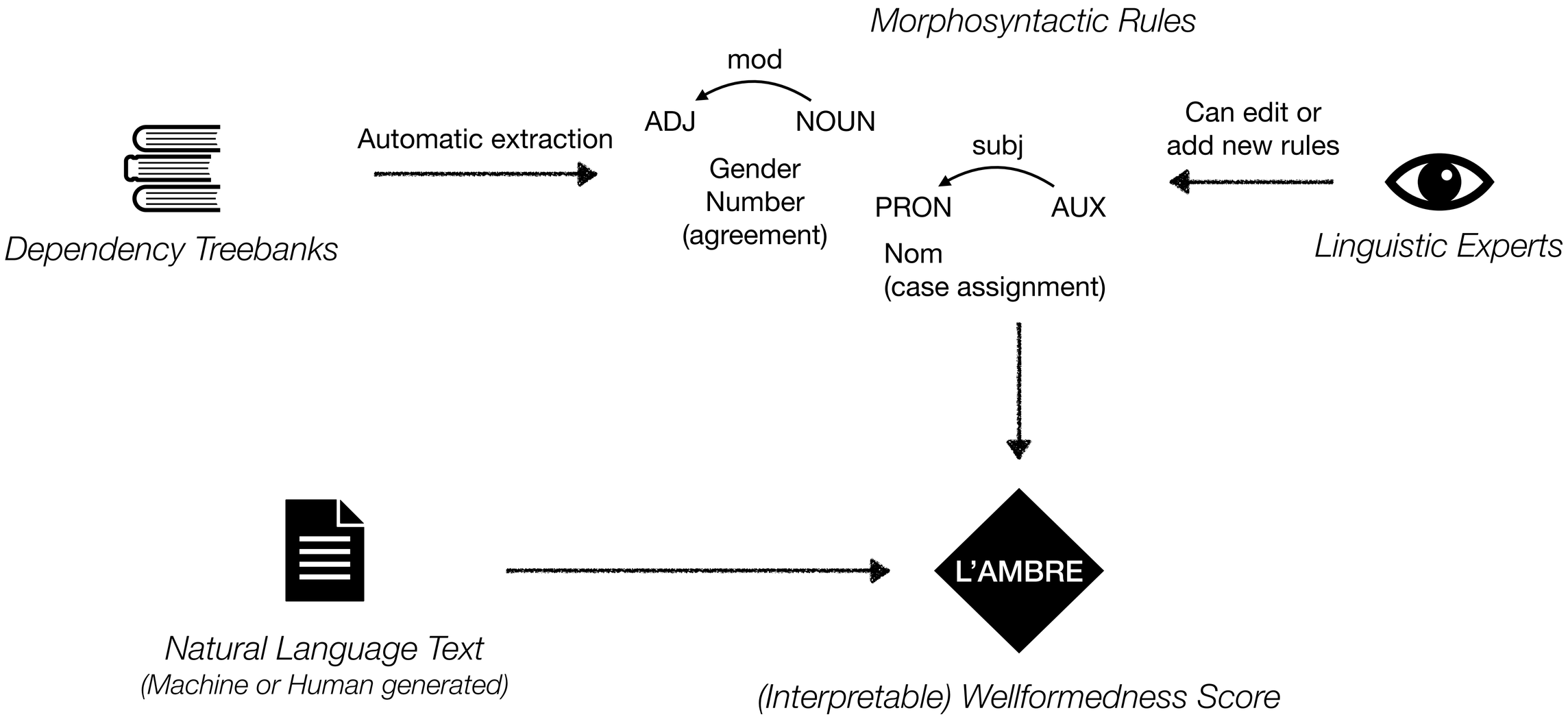
- Wellformedness is usually necessary for fluency but not sufficient (Sakaguchi et al., 2016)
- We focus on wellformedness for two reasons,
  1. Easier to measure via a simple comparison against a checklist
  2. Wellformedness rules can be extracted automatically and accurately

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. *Reassessing the goals of grammatical error correction: Fluency instead of grammaticality*. TACL 2016.

New Metric: L’AMBRE

- **L**inguistically **A**ware **M**orphosyntax-**B**ased **R**ule **E**valuation
- Evaluates natural language text by checking against language-specific morphosyntactic rules
- Metric is interpretable by nature
- Morphosyntactic rules are automatically extracted from dependency treebanks, can be easily extended to new languages

L’AMBRE Overview



Morphosyntactic Rules

- We focus on three types of rules,
  - Agreement
  - Case assignment
  - Verb form choice
- Extract directly from SUD treebanks (Chaudhary et al. 2020)

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. *Automatic extraction of rules governing morphological agreement*. EMNLP 2020.

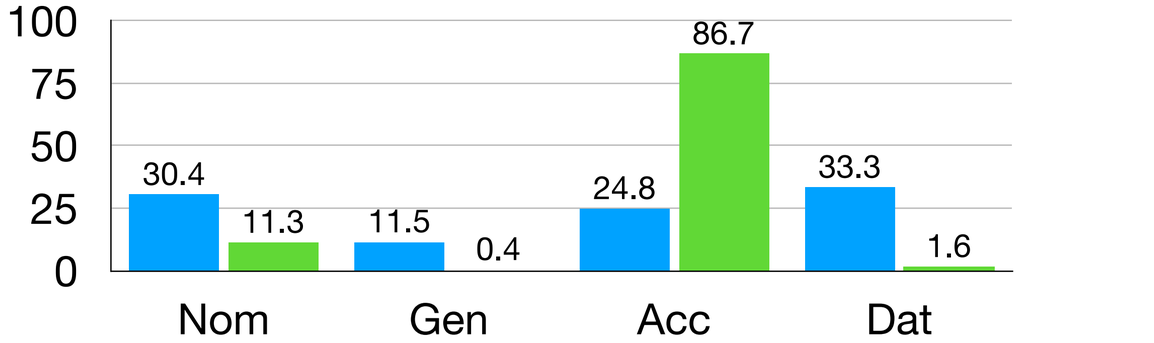
Extracting Rules

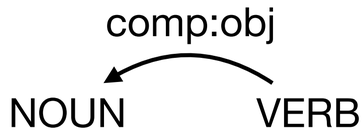
Agreement

$$\frac{\#\,[\text{Gender}_{ADJ} = \text{Gender}_{NOUN}]}{\#\,\left[\begin{array}{c}\text{mod}\\ \text{ADJ} \quad \text{NOUN}\end{array}\right]} >= 0.9$$

Gender
(agreement)

Case Assignment

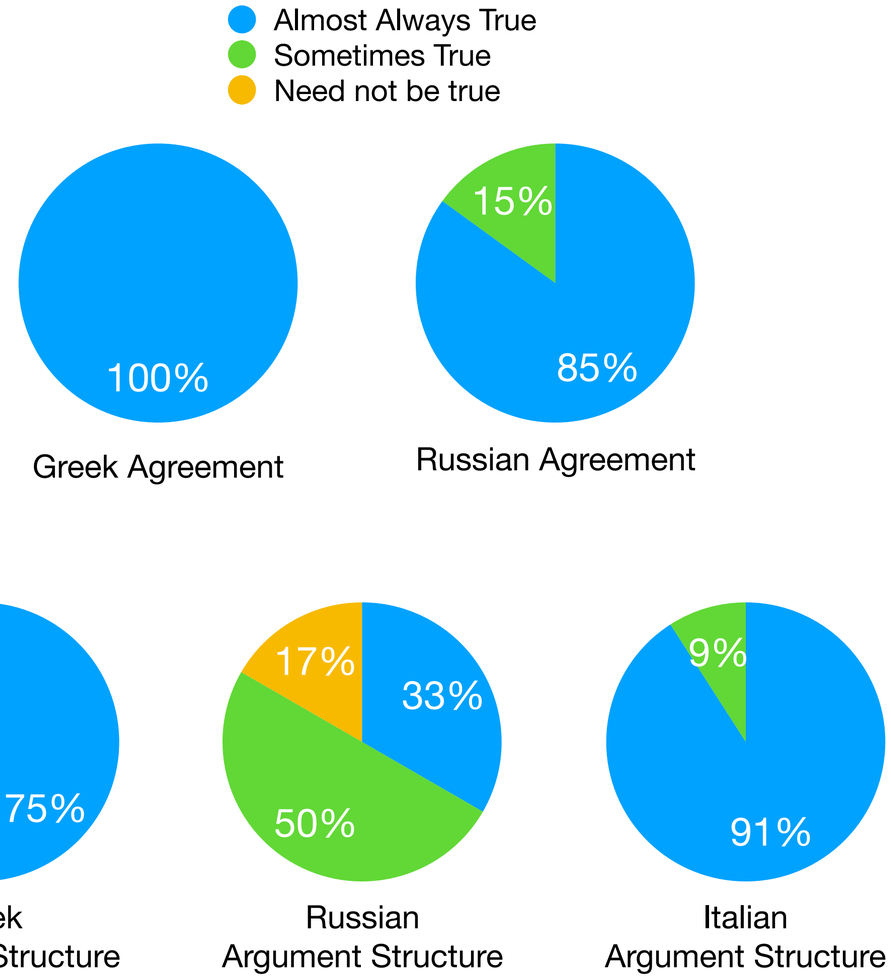




Acc
Nom
(case assignment)

Human Evaluation of Rules

- We ask human experts to label each rule as one of
  - almost always true
  - sometimes true
  - need not be true
- Our rules are of high quality



Checking Rules

- **L'AMBRE**: parse text and match against the morphosyntactic rules
- But…
  - Noisy (or grammatical incorrect) text can throw off the standard parsers (Hashemi and Hwa, 2016 and others)

Homa B. Hashemi and Rebecca Hwa. *An evaluation of parser robustness for ungrammatical sentences*. EMNLP 2016.

Making Parsers Robust

- Utilize UniMorph to add morphology related noise to standard treebanks
- Retrain Stanza parsers on noisy treebanks
- Improved results in feature tagging and dependency parsing on noisy text

Treebank Sentence

(οικισμός + N;ACC;SG)

Στο	μικρό	**οικισμό**	της	Λίνδου
In-the	small	settlement	of	Lindos

Στο	μικρό	**οικισμούς**	της	Λίνδου

(οικισμός + N;ACC;PL)

Noisy Sentence

Validating L’AMBRE: Grammatical Error Correction (GEC)

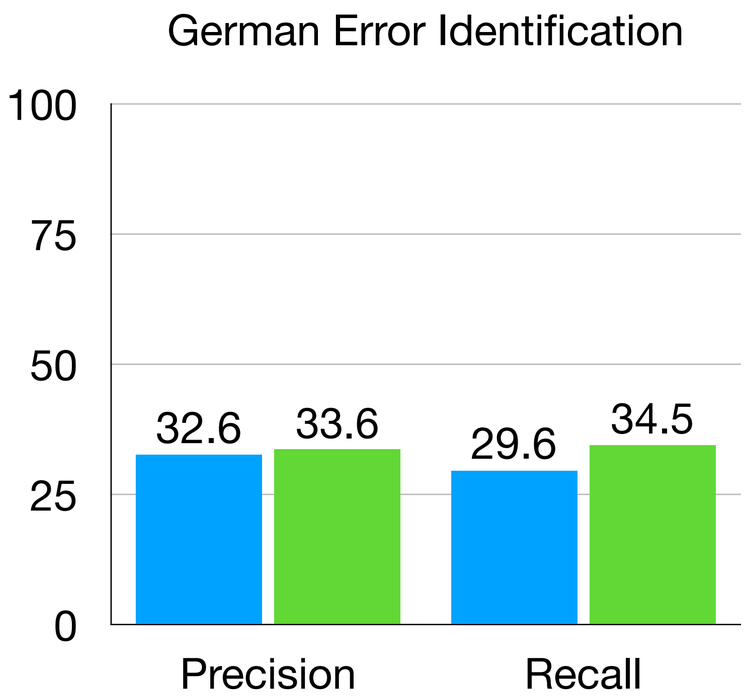
- Does L’AMBRE accurately capture grammaticality?
- Validate on Russian and German GEC
- We focus on error identification (not correction)

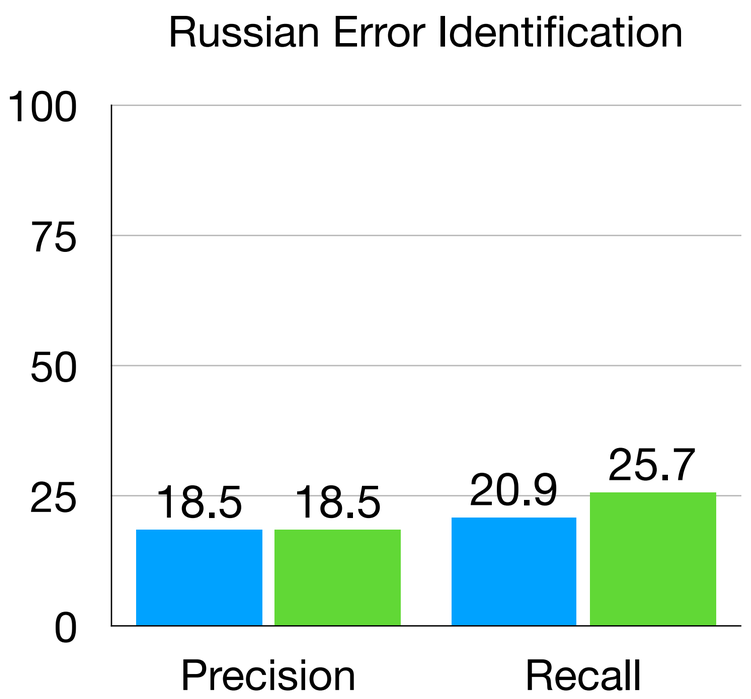
			ihr	aktuelles	
Vielleicht	verdienen	Ärzte	~~ihre~~	~~aktuelle~~	Gehalt
Maybe	*to earn*	*doctors*	*their*	*current*	*salary*

Example from German GEC dataset

Results

- Precision and Recall of detecting morphology related errors
- Robust parsers improve on recall





Results

- Comparison to other metrics for error identification
  - grammaticality-based metric (Napoles et al. 2016; GBM)
  - LM perplexity
- Observations
  - Perplexity performs the best but not interpretable
  - L'AMBRE does well at morphology related errors, and GBM does well with other fluency related rules

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. *There's no comparison: Referenceless evaluation metrics in grammatical error correction*. EMNLP 2016.

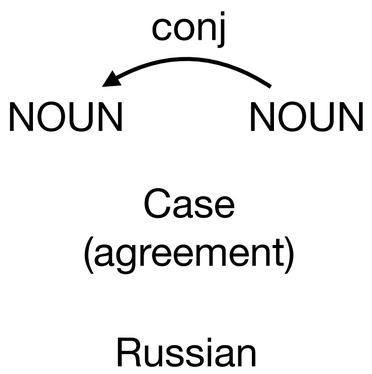
NMT Case Study

- Analyze systems from WMT shared tasks
- L'AMBRE correlates well with human scores
- L'AMBRE also correlates well with perplexity
- L'AMBRE can provide fine-grained analysis

English → X	WMT '18	WMT '19
Czech	0.84	0.80
German	-0.06	0.16
Estonian	0.68	-
Finnish	0.86	0.85
Russian	0.86	0.57
Turkish	0.58	-

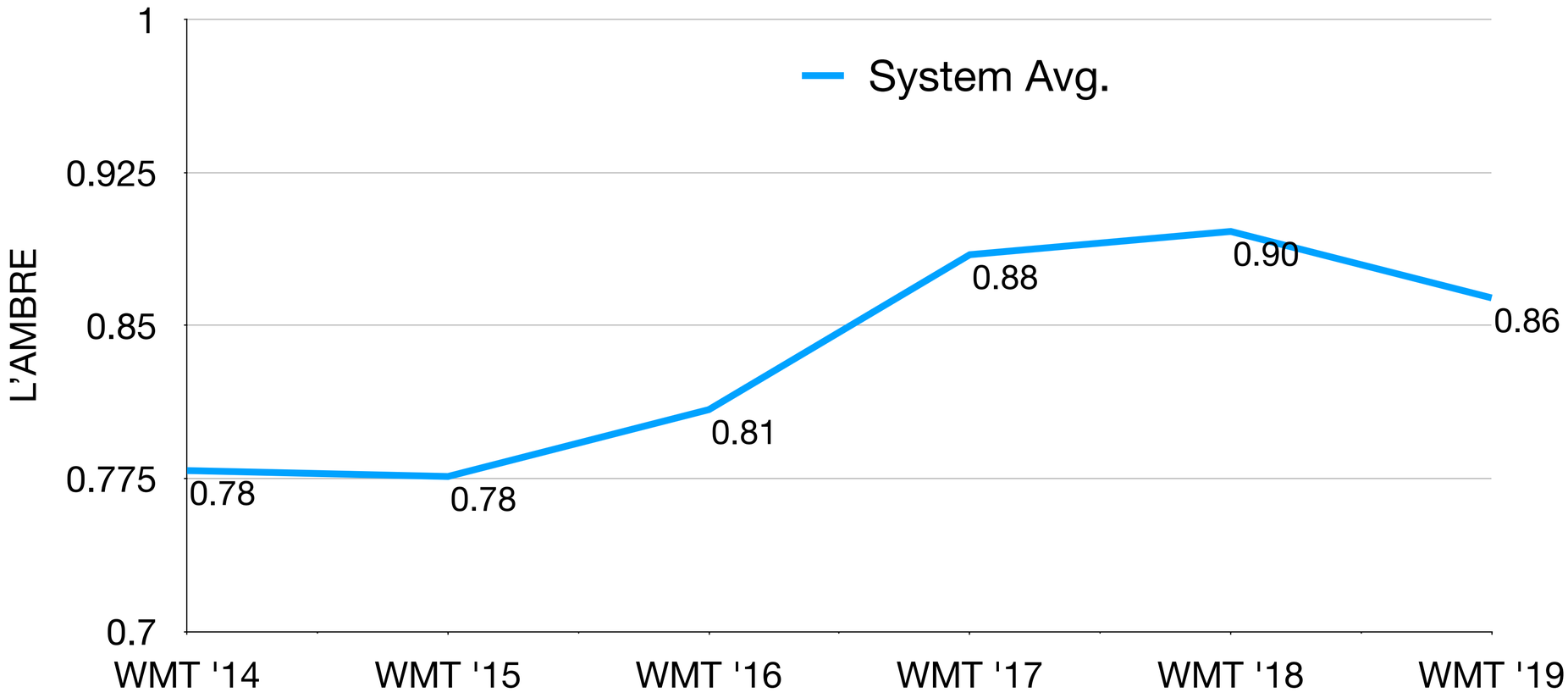
Pearson's r Correlations to human z-scores

Fine-grained Diachronic Analysis

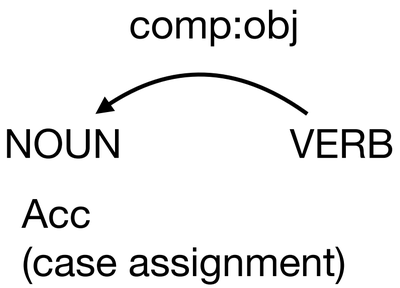


Case
(agreement)

Russian

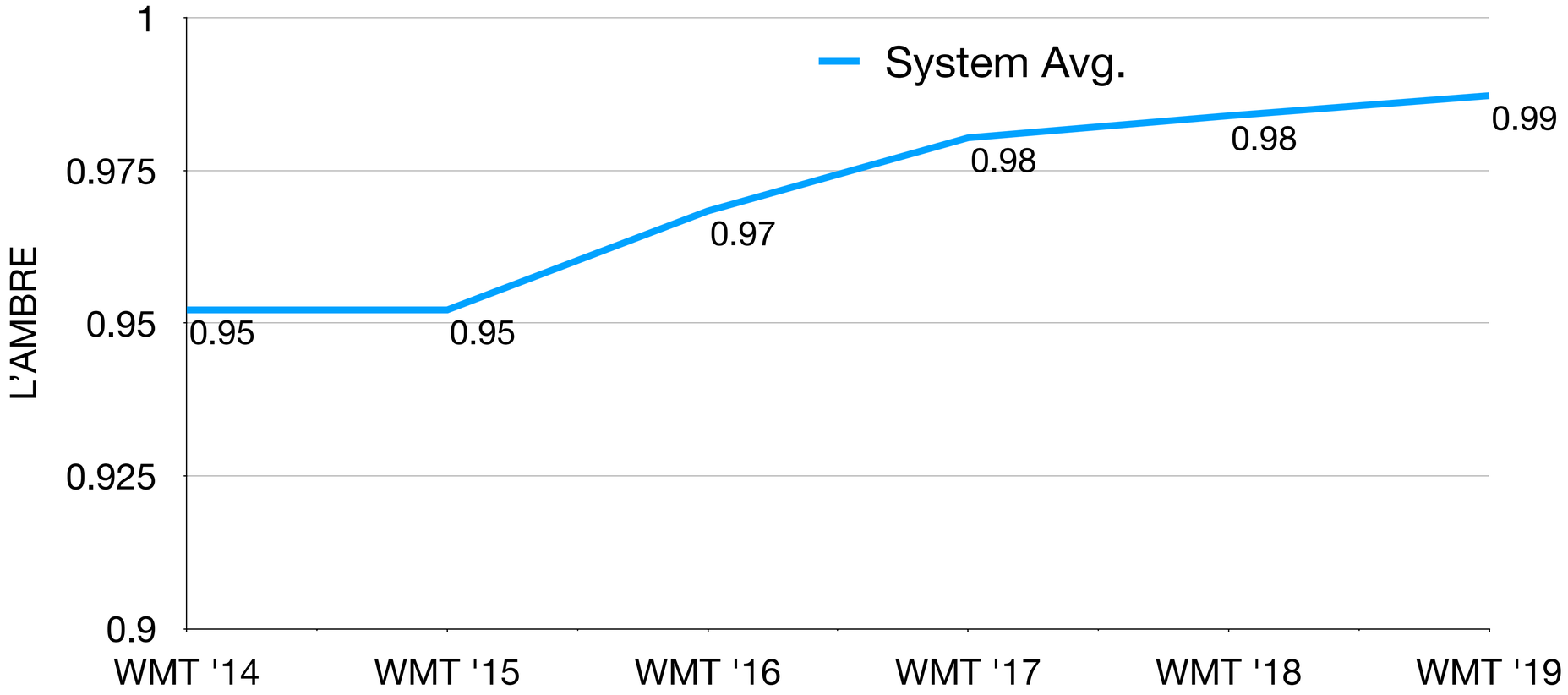


Fine-grained Diachronic Analysis



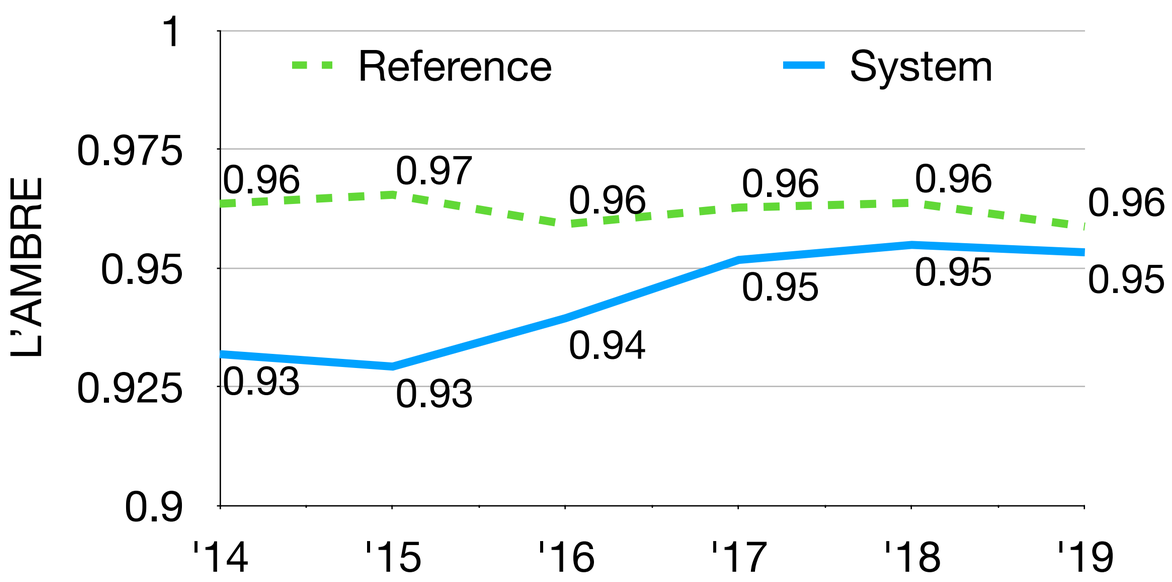
Acc
(case assignment)

Russian

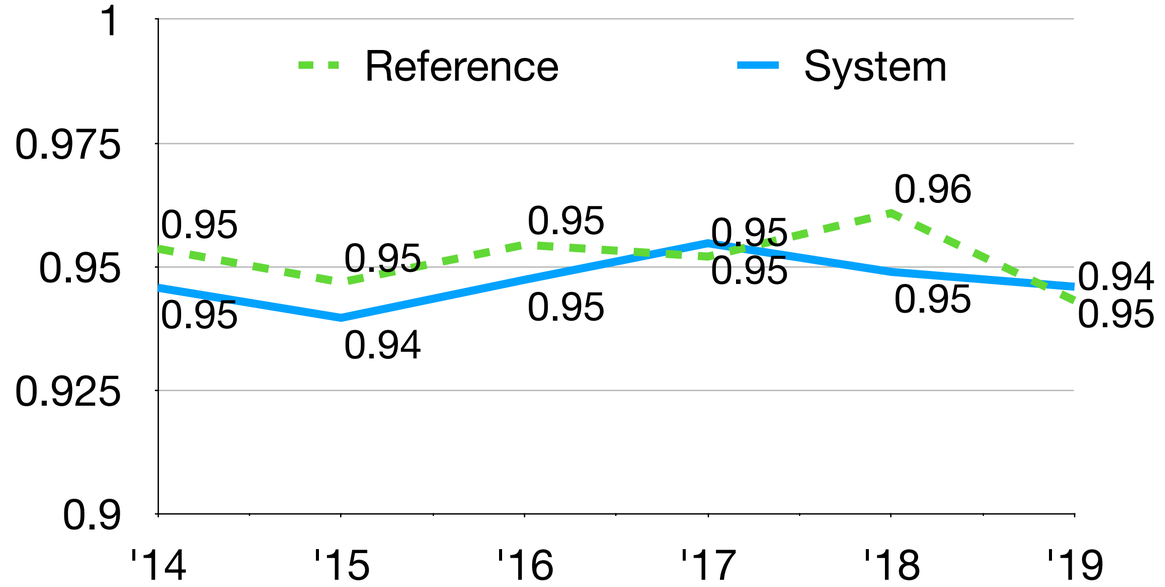


Diachronic Analysis

Russian WMT



German WMT



Summary

- Our contributions,
  - **L'AMBRE**: a multilingual metric to measure wellformedness
  - **Robust parsers**: make parsers and taggers robust by adding morphological noise
  - **Case Study on NMT**: applications in evaluating WMT system outputs

github.com/adithya7/lambre