Dutch Historical Spelling Normalization for Parsing and Coreference Resolution

Priscilla Postma, Rina Donker, Ruth Stam, Athalia Roorda,
Andreas van Cranenburgh, Gertjan van Noord
CLCG, University of Groningen

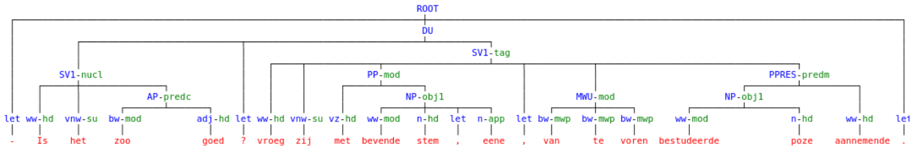
Non-canonical language can be handled in an NLP pipeline using normalization of the input (e.g., MoNoise; van der Goot & van Noord, 2017) or domain adaptation of the pipeline (e.g., Hupkes & Bod, 2016); we focus on the former. MoNoise shows that normalization is effective for social media language. We consider a different domain: Dutch literature from Project Gutenberg. We work with 9 fragments that make up the OpenBoek corpus (van den Berg et al., 2021). The fragments consist of 10,000+ tokens from texts first published 1860-1920, both translated and originally Dutch.

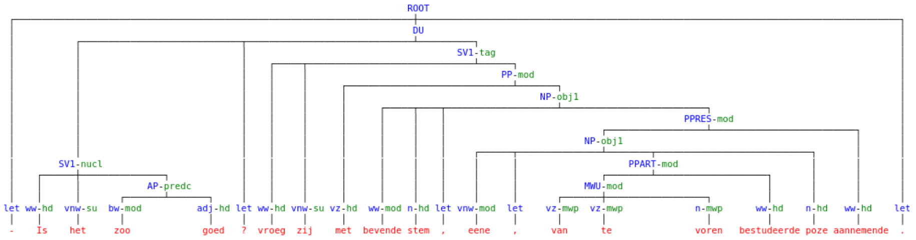
MoNoise consists of several modules: a lookup table, automatic spelling correction (`aspell`), and word embeddings; we aim to explore these techniques on our data in future work. Here we report results of a rule-based approach implemented with a `sed` script (i.e., regular expressions) for normalizing frequently occurring non-standard spellings.

The output consists of instructions to the Alpino parser (van Noord, 2006) to treat words with non-canonical orthography as if they occur with modern spelling. The advantage of this approach is that the resulting parse trees contain the original tokens, and existing annotation layers (such as coreference) do not have to be re-aligned. Consider the following sentence from Couperus, *Eline Vere* (ch. I, § II):

```
18-1|- Is het [ @alt zo zoo ] goed ? vroeg zij met
bevende stem , [ @alt ene eene ] , van te voren
bestudeerde poze aannemende .
```

Here `[ @alt zo zoo ]` indicates that the original token *zoo* should be treated as *zo*. Besides doubled vowels, other frequent spelling normalizations are *de/den*, *zei/zeide*, and *mensen/menschen*. When multiple alternatives are given the parser considers the input as a lattice and uses the sequence of tokens that generates the most likely parse. Parse trees for the above sentence (original, normalized):


ROOT
DU
SV1-tag
SV1-nucl
PP-mod
PPRES-predm
AP-predc
NP-obj1
MWU-mod
NP-obj1
let ww-hd vnw-su bw-mod adj-hd let ww-hd vnw-su vz-hd ww-mod n-hd let n-app let bw-mwp bw-mwp bw-mwp ww-mod n-hd ww-hd let
- Is het zoo goed ? vroeg zij met bevende stem , eene , van te voren bestudeerde poze aannemende .


ROOT
DU
SV1-tag
PP-mod
NP-obj1
PPRES-mod
NP-obj1
SV1-nucl
PPART-mod
AP-predc
MWU-mod
let ww-hd vnw-su bw-mod adj-hd let ww-hd vnw-su vz-hd ww-mod n-hd let vnw-mod let vz-mwp vz-mwp n-mwp ww-hd n-hd ww-hd let
- Is het zoo goed ? vroeg zij met bevende stem , eene , van te voren bestudeerde poze aannemende .

While the automatic spelling normalization is not perfect (the correct normalization of *eene* is *een* with POS `lid` rather than *ene*), it does lead to a correct bracketing of the NP *eene* … *poze*. Furthermore, it turns out that a comma is missing after *bestudeerde* in the Project Gutenberg etext we use (EBook-No. 19563); the DBNL version of this text (`coup002elin01_01`) does have this comma—this underscores the importance of professionally edited critical editions.

We will perform an intrinsic evaluation of our spelling normalization pipeline with manually corrected texts and report F1 scores (Reynaert, 2008). We also perform an extrinsic evaluation of downstream tasks: part-of-speech tagging, mention detection, and coreference resolution. Scores for the latter two tasks on Multatuli, *Max Havelaar*:

	mentions			lea				pron
	recall	prec	f1	recall	prec	f1	CoNLL	acc
original	89.96	81.29	85.40	54.80	47.07	50.64	65.76	55.00
normalized	90.18	82.22	86.02	54.82	45.96	50.00	65.48	54.20

The mention score is improved, which makes sense given that parsing of NPs seems to improve after spelling normalization, but there is a decrease in the coreference metrics, which warrants further investigation.

References

Frank van den Berg, Esther Ploeger, Menno Robben, Pauline Schomaker, Robin Snoek, Remi Thüss, Andreas van Cranenburgh (2021). OpenBoek: A Corpus of Coreference and Entities in Dutch Literature. CLIN 31. https://andreasvc.github.io/openboek/clin2021abstract.pdf

Rob van der Goot, Gertjan van Noord (2017). MoNoise: Modeling Noise Using a Modular Normalization System. CLIN Journal, 7, 129–144. https://www.clinjournal.org/index.php/clinj/article/view/74

Dieuwke Hupkes, Rens Bod (2016). POS-tagging of Historical Dutch. Proceedings of LREC. https://aclanthology.org/L16-1012/

Gertjan van Noord (2006). At Last Parsing Is Now Operational. Proceedings of TALN. http://www.let.rug.nl/vannoord/papers/taln.pdf

Martin Reynaert (2008). All, and only, the errors: more complete and consistent spelling and OCR-error correction evaluation. Proceedings of LREC. https://aclanthology.org/L08-1217/