

Software Project NLP with NNs

# Proto-Word Reconstruction with RNNs

## Project Sketch

Julius Steuer   Morgan Wixted

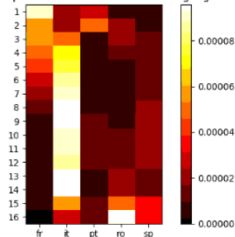
August 6, 2020

# Baseline

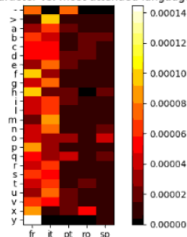
Start from Meloni et al. (2019):

- Romance dataset (not publicly available)
- Character-based encoder-decoder architecture with Bahdanau attention
- Characters encoded as 100-bit vectors (localist representation), character embeddings shared over all languages
- Output one character of the reconstructed word per timestep
- Difference to Machine Translation:
  - Sentence  $\rightarrow$  word, word  $\rightarrow$  character
- Evaluate impact of individual languages on reconstruction:

Output positions vs. most-attended language



Output character vs. most-attended language



# Model

TODO: Feature encoding

Idea:

- Use distributed feature encodings to represent characters, not embeddings
- Use the stem of the Latin word as the proto-form, since nominative inflection is seldom preserved in modern Romance
- Start with ASJP (easily available) transcriptions, later switch to IPA (richer annotation).
- Pairwise alignments of sequence chunks (following Ciobanu and Dinu (2018))

Lat.	acc.	1	2	3	4	5	6	7	8	9	10
<hr/>											
lat.	-	k	o	-	r	t	i	k	-	e	
it.	-	k	o	-	r	t	e	t	C	a	
spa.	-	k	o	-	r	t	e	8	-	a	
fr.	e	k	o	-	r	-	-	-	-	-	
rom.	s	k	o	a	r	-	-	c	-	3	

## Model, continued

Input:

- One column is represents the input at a single timestep:

$$I_{T=t_1} = (-, -, -, e, s)$$

$$I_{T=t_2} = (k, k, k, k, k)$$

...

$$I_{T=t_{10}} = (e, a, a, -, 3)$$

- Attention: Meloni et al. (2019) attend on different languages in the cognate set.

In our case it depends on how we present the data to the model:

Either (as in the paper) perform a pass through the model to reconstruct a single character (each  $I_{T=t_i}$  a sequence of inputs), or

- Use the matrix/vector  $I_{T=t_i}$  at a single time step, and perform only one pass

TODO: Draw model architecture

# Model, expectations

But: The direct precursor of it. *corteccia*, spa. *corteza* etc. is the adjective lat. *corticeus*, -a, -um

- We expect the model to reconstruct an a-stem noun instead of a consonant stem
- Spurious sounds (fr. *e-*, rom. *s-*) should be dropped
- Ambiguous sounds may not indicate a clear rule (lat. *-k-* vs. rom. *-c-* vs. spa. *-8-*)

TODO: more items

# Milestones

Minimal:

- Reconstruction with ASJP encodings, small Swadesh list as data
- Then switch to IPA encodings
- Try different language family (initial guess on unseen data)
- Examine the influence of different (or absent) alignments on model performance

Great to have:

- Use larger dataset (asked Ciobanu for hers)
- Ensure compatibility with the LingPy <sup>1</sup> WordList <sup>2</sup> class

---

<sup>1</sup><http://lingpy.org/>

<sup>2</sup><http://lingpy.org/reference/lingpy.basic.html#lingpy.basic.wordlist.Wordlist>

Timeline?

# References

- Ciobanu, A. M. and Dinu, L. P. (2018). Ab Initio: Automatic Latin Proto-word Reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Meloni, C., Ravfogel, S., and Goldberg, Y. (2019). Ab Antiquo: Proto-language Reconstruction with RNNs. arXiv: 1908.02477.