Language-agnostic measures discriminate inflection and derivation

COLEMAN HALEY   EDOARDO PONTI   SHARON GOLDWATER

THE UNIVERSITY *of* EDINBURGH
UKRI Centre for Doctoral Training
in Natural Language Processing

THE UNIVERSITY *of* EDINBURGH
informatics

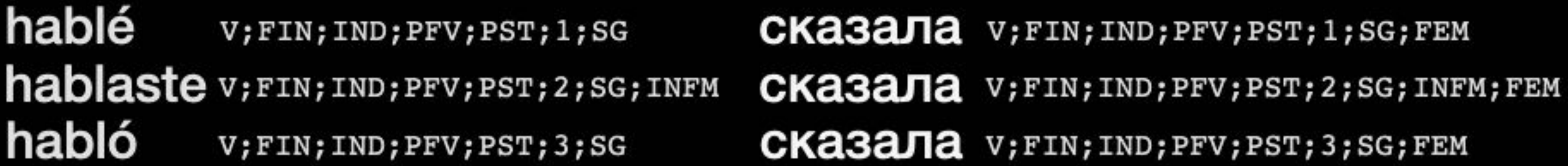
UK Research
and Innovation

What is the difference between *constructed* and *constructor*?

INFLECTION

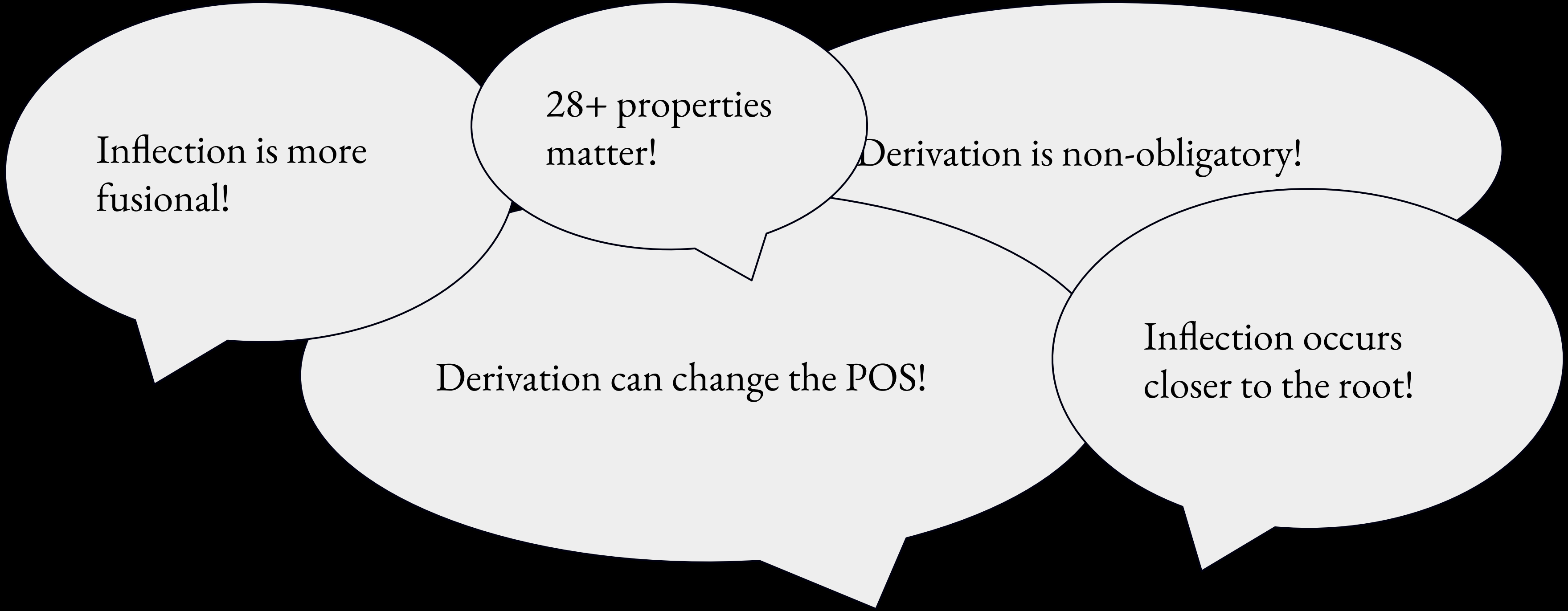
DERIVATION

Apparent cross-linguistic agreement about what *forms* are inflectional



hablé	V;FIN;IND;PFV;PST;1;SG	сказала	V;FIN;IND;PFV;PST;1;SG;FEM
hablaste	V;FIN;IND;PFV;PST;2;SG;INFM	сказала	V;FIN;IND;PFV;PST;2;SG;INFM;FEM
habló	V;FIN;IND;PFV;PST;3;SG	сказала	V;FIN;IND;PFV;PST;3;SG;FEM

But what properties *underly* this distinction?



Inflection is more fusional!

28+ properties matter!

Derivation is non-obligatory!

Derivation can change the POS!

Inflection occurs closer to the root!

Can such properties describe the distinction across a wide range of languages?

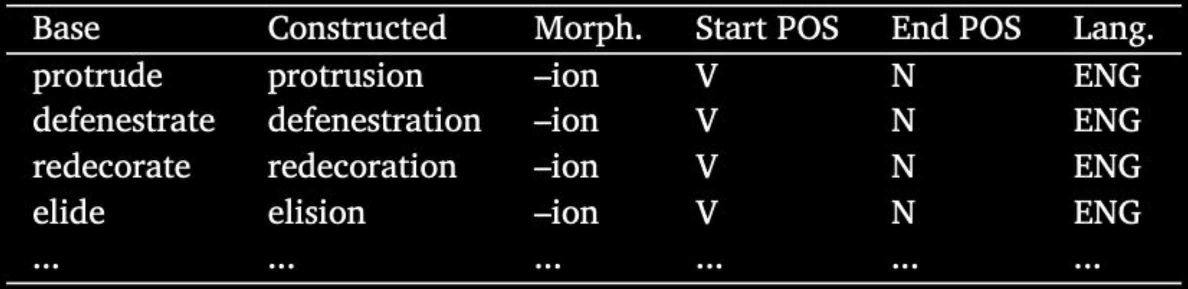
Goal: find corpus-based measures that can discriminate inflection and derivation in UniMorph

Intuition: derivational constructions produce *larger* and more *variable* changes to words

- in terms of form (edit distance)
- in terms of distribution (FastText embedding)

	magnitude	variability
form	$\|\|\Delta_{\text{form}}\|\|$	$\text{var}(\Delta_{\text{form}})$
distribution	$\|\|\Delta_{\text{embed}}\|\|$	$\text{var}(\Delta_{\text{embed}})$

We look at constructions across 26 languages

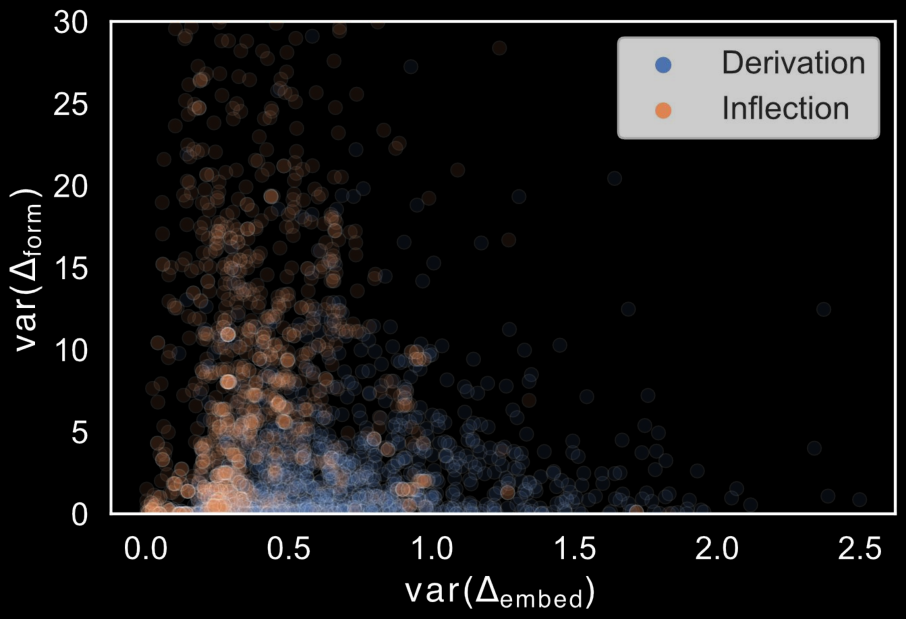


Base	Constructed	Morph.	Start POS	End POS	Lang.
protrude	protrusion	–ion	V	N	ENG
defenestrate	defenestration	–ion	V	N	ENG
redecorate	redecoration	–ion	V	N	ENG
elide	elision	–ion	V	N	ENG
...	...	...	...	...	...

Unimorph assigns inflection & derivation consistently in terms of these features!

1. We train a classifier with these 4 features and predict on held-out constructions
2. Majority-class baseline is 57% accuracy
3. 86% accuracy using all features in a linear classifier, 90% accuracy using an MLP
4. No language-specific features!
5. Seems to generalize to non-Indo-European languages

Consistent, but gradient:



Conclusions

1. The inflection-derivation distinction can be recovered from corpora with accuracy as high as 90%
2. But the distinction appears gradient

Reach me at: coleman.c.haley@gmail.com

@colemanhaley22 on Twitter