# Cross-Lingual Morphological Tagging for Low-Resource Languages

Jan Buys* Jan A. Botha†

University of Oxford*, Google Inc†

## Summary

**Goal**: Train morphological taggers for languages without annotated training data.

**Approach**

- Constrain candidate tags using word-aligned parallel text.
- Perform scalable, weakly-supervised learning with Wsabie.

**Outcome**

- Promising intrinsic performance, depending on language-pair choice.
- Induced taggers improve downstream dependency parsing.

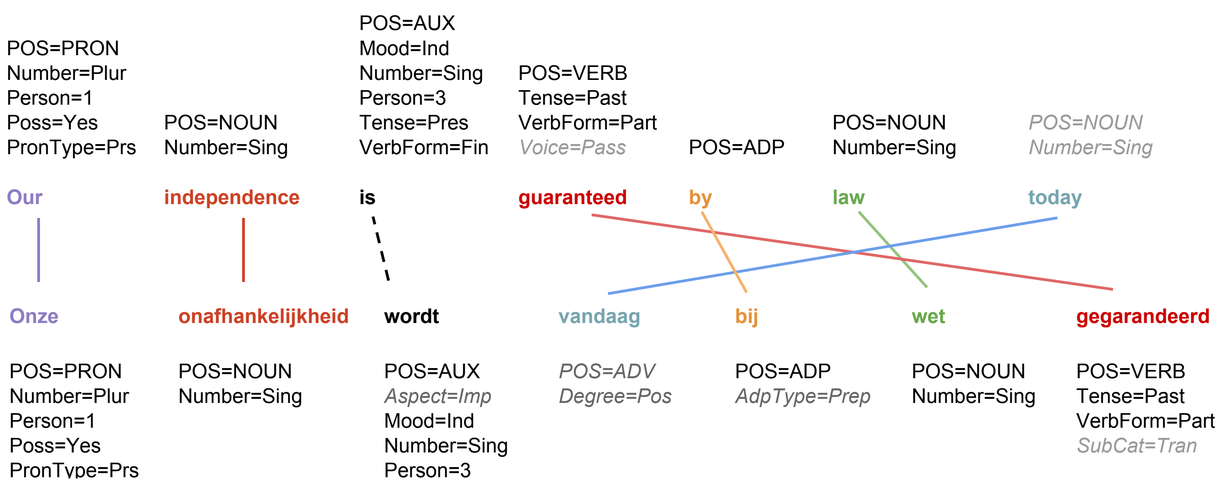
## Projecting Universal Morphological Tags

Projection from high-resource to low-resource languages for *PoS taggers* has been shown to work. We extend this to ***morphological taggers***, which have much larger **tag sets featuring hundreds of tags**.

Universal morphological tags are sets of attribute-value pairs annotated in the Universal Dependency (UD) Treebanks – consistent across languages, but language-specific phenomena cause partial mismatches.

Using word-aligned bitext, we extract a tag dictionary for the target-side word types and, optionally, constrain a target token's tag to that of its aligned source token. A tagger is trained on the target-side text, subject to one or both kinds of constraints.

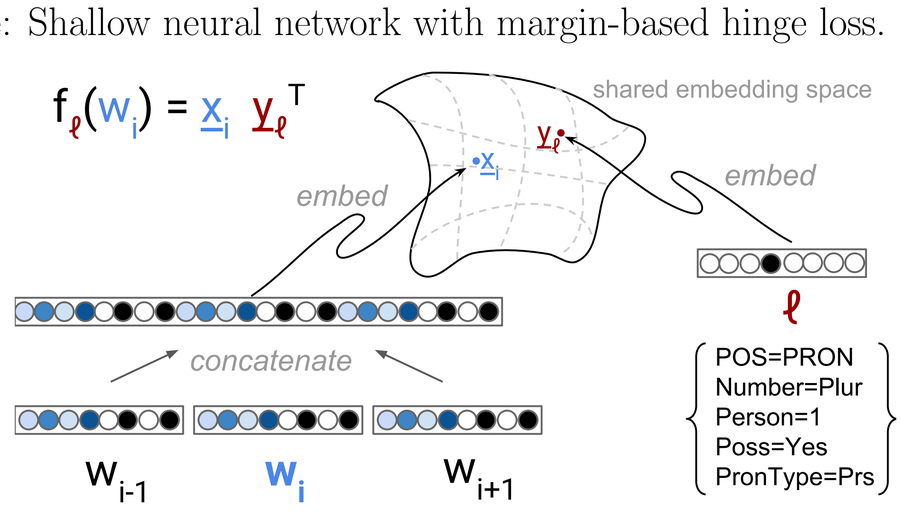
## Example




A parallel sentence in English and Dutch annotated with universal morphological tags, showing high-confidence automatic word-alignments.

## Models

**Baseline**: Generative HMM with log-linear parameterization.
**Wsabie**: Shallow neural network with margin-based hinge loss.




- Learns to rank tags licensed by projected constraints above all other tags.
- Runtime is linear in number of tags, in contrast to quadratic Viterbi.

## Model Comparison

| | |
|---|---|
| HMM with projected tag dictionary | 53.86 |
| HMM with projected tag dictionary and token-constraints | 48.49 |
| projected tag dictionary, unambiguous word types only | 51.72 |
| projected tag dictionary | 53.60 |
| projected tag dictionary and token-constraints | 53.36 |
| fully supervised model trained on 1000 tokens | 62.44 |
| oracle tag dictionary | 75.55 |

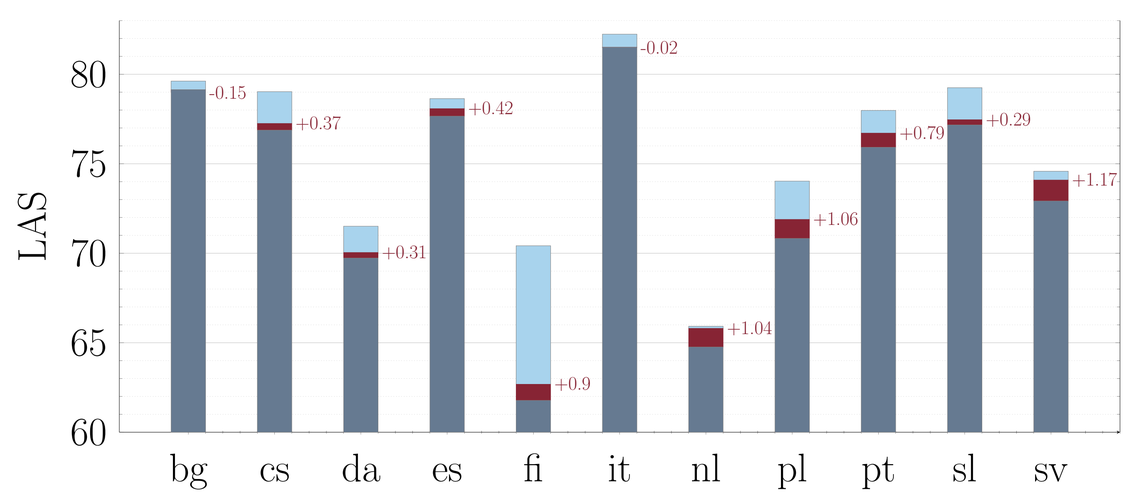
Cross-lingual morphological tagging in 11 languages that have both UD treebanks and Europarl bitext. Macro-averaged F1 scores, with English as source language. All but the first two rows are Wsabie models.

## Results: Cross-lingual Tagging

| ↓src/trg: | bg | cs | da | es | fi | it | nl | pl | pt | sl | sv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| en | 46.7 | 49.7 | 58.0 | 55.7 | 54.0 | 59.6 | **64.1** | 45.0 | 57.8 | 51.0 | 47.9 |
| bg | - | 58.3 | 59.2 | 51.2 | 52.6 | 43.2 | 38.7 | 52.8 | 41.1 | 49.2 | 53.6 |
| cs | 55.2 | - | 54.5 | 42.3 | 48.4 | 51.3 | 45.0 | **56.8** | 33.6 | **67.5** | 53.2 |
| da | **61.9** | **61.6** | - | 41.8 | 49.1 | 45.5 | 49.6 | 53.7 | 44.0 | 49.3 | **72.1** |
| es | 54.3 | 58.8 | 41.3 | - | 53.0 | **74.4** | 52.1 | 52.2 | **69.2** | 53.8 | 46.9 |
| fi | 46.6 | 48.7 | 45.3 | 39.5 | - | 50.9 | 36.8 | 37.4 | 30.1 | 55.5 | 57.8 |
| it | 43.6 | 59.4 | 44.0 | **74.0** | 53.3 | - | 54.3 | 46.5 | 69.2 | 55.9 | 47.0 |
| nl | 44.7 | 59.5 | 56.2 | 54.8 | 54.0 | 60.3 | - | 55.9 | 58.6 | 48.6 | 51.6 |
| pl | 52.7 | 58.6 | 46.3 | 37.5 | 42.1 | 47.9 | 42.1 | - | 40.7 | 56.0 | 42.6 |
| pt | 45.4 | 45.0 | 49.6 | 66.2 | 42.6 | 69.5 | 50.1 | 43.5 | - | 47.8 | 43.9 |
| sl | 46.6 | 60.7 | 35.2 | 40.9 | 49.2 | 49.8 | 36.0 | 54.1 | 35.0 | - | 40.4 |
| sv | 50.1 | 54.6 | **70.7** | 47.7 | **57.2** | 49.7 | 46.9 | 41.6 | 46.3 | 43.5 | - |

Macro-averaged F1 scores for all language pairings, using Wsabie model with projected tag dictionary.

## Results: Dependency Parsing




- Supervised morphological tagger (trained using full treebank)
- Projected tagger (Wsabie with projected tag dictionary, source English)
- Baseline — no morphology

Dependency parsers trained with gold PoS and parse trees, but varying the origin of morphological annotations used as parser features.
*Avg. LAS gain over baseline:* Projected tagger +0.6; Supervised tagger +2.3.

## Future Work

- Strategies for choosing (or combining multiple) source languages.
- Leverage Wiktionary for constraints.
- Incorporate source-side syntactic information.