

# Robust Morphological Tagging with Word Representations

**Thomas Müller** and **Hinrich Schütze**  
Center for Information and Language Processing  
University of Munich, Germany  
muellets@cis.lmu.de

|       |    | word |     | word-tag |     |
|-------|----|------|-----|----------|-----|
|       |    | ID   | OOD | ID       | OOD |
| morph | cs | .09  | .13 | .03      | .16 |
|       | de | .08  | .13 | .04      | .06 |
|       | en | .03  | .10 | .00      | .02 |
|       | es | .65  | .13 | .00      | .02 |
|       | hu | .20  | .27 | .00      | .02 |
|       | la | .18  | .37 | .02      | .04 |

Table 1: Rates of unknown words and word-tag combinations in ID and OOD development sets

|    | POS | MORPH | train   | ID     | OOD    |
|----|-----|-------|---------|--------|--------|
| cs | 12  | 450   | 652,544 | 87,988 | 27,350 |
| de | 54  | 681   | 719,530 | 76,704 | 24,622 |
| en | 48  | 48    | 731,678 | 32,092 | 53,156 |
| es | 12  | 288   | 427,442 | 50,368 | 56,638 |
| hu | 22  | 673   | 170,141 | 29,989 | 83,087 |
| la | 23  | 749   | 59,992  | 9,475  | 41,432 |

Table 2: Labeled data set statistics. Number of part of speech tags (POS) and morphological tags (MORPH); number of tokens in training set (train), ID development set and OOD development set

|    | articles  | tokens        | types     |
|----|-----------|---------------|-----------|
| cs | 270,625   | 93,515,197    | 1,607,183 |
| de | 1,568,644 | 682,311,227   | 7,838,705 |
| en | 4,335,341 | 1,957,524,862 | 7,174,661 |
| es | 1,004,776 | 432,596,475   | 6,033,105 |
| hu | 245,558   | 95,305,736    | 2,776,681 |
| la | 5,316     | 88,636,268    | 713,162   |

Table 3: Number of articles, tokens and types in the unlabeled data sets

|    | ID  |      | OOD  |      |
|----|-----|------|------|------|
|    | ALL | OOV  | ALL  | OOV  |
| cs | 4.7 | 42.8 | 6.5  | 45.6 |
| de | 7.7 | 55.0 | 8.4  | 50.6 |
| en | 0.9 | 23.8 | 2.1  | 22.7 |
| es | 5.5 | 37.5 | 5.4  | 29.5 |
| hu | 9.9 | 37.6 | 11.3 | 38.0 |
| la | 2.0 | 8.0  | 6.8  | 17.6 |

Table 4: Percentage of tokens not covered by the representation vocabulary

|           |     | Domain / Source  | Reference                                  |
|-----------|-----|--|--|
| English   | ID  | news text  | (Petrov et al., 2012)                      |
|           | OOD | Yahoo! Answers, weblogs, news groups, buisness reviews, emails | (Petrov et al., 2012)                      |
| Czech     | ID  | news text  | (Böhmová et al., 2003; Hajič et al., 2009) |
|           | OOD | novel  | (Erjavec, 2010)                            |
| German    | ID  | news text  | (Brants et al., 2002)                      |
|           | OOD | hiking (Swiss German), DVD player manual, novel, economics     | (Volk et al., 2010)                        |
| Hungarian | ID  | news text  | (Vincze et al., 2010; Seddah et al., 2013) |
|           | OOD | Windows 2000 manual, novel                                     | (Vincze et al., 2010)                      |
| Latin     | ID  | bible text ( <i>Vulgate</i> )                                  | (Haug and Jøhndal, 2008)                   |
|           | OOD | <i>Bello Gallico, Cicero, Aetheria</i>                         | (Haug and Jøhndal, 2008)                   |
| Spanish   | ID  | news text  | (Taulé et al., 2008; Hajič et al., 2009)   |
|           | OOD | law, economics, medicine, computer science, enviroment         | (Marimon et al., 2012)                     |

Table 5: Domains and data sources for each language.