Appendix: How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?

Hila Gonen[1] **Yova Kementchedjhieva**[2] **Yoav Goldberg**[1,3]
[1]Department of Computer Science, Bar-Ilan University
[2]University of Copenhagen
[3]Allen Institute for Artificial Intelligence
hilagnn@gmail.com,yova@di.ku.dk,yoav.goldberg@gmail.com

A Implementation Details

Morphological Analyzers For Italian, we use Morph-it!,[1] a lexicon of inflected forms with their lemma and morphological features. For German, we use DEMorphy,[2] which, given a word, provides its full morphological analysis (or several, when applicable) (Altinok, 2018).[3]

Training Word Embeddings We train 300d word embeddings with window size 4, on January 2018 wikipedia dump[4] for all three languages. After tokenization we get 2.2B (En), 463M (It) and 815M (De) tokens. We discard words that do not appear at least 50 times, and are left with vocabulary sizes of 360,386 (En), 161,144 (It) and 361,944 (De). We train using word2vecf (Levy and Goldberg, 2014), which allows to change context words without affecting target words.

B Manual Mapping for Italian

Tables 1 and 2 contain the manual mappings we used for Italian (see next page).

C Qualitative Evaluation

As a qualitative evaluation, we take several words for SimLex-999 and look at their top-10 nearest neighbor lists, before and after applying our method. In Table 3 we show the top-10 lists for the words *palla* (ball-feminine) in Italian, and *diamant* (diamond-masculine) in German. It is evident that the words that are added to the list are better correlated with the target word than those that are removed (see next page).

[1]http://tools.sslmit.unibo.it/doku.php?id=resources:morph-it

[2]https://github.com/DuyguA/DEMorphy

[3]Since the analysis is fine-grained, when searching for a different-gender word form, we do not require full match, but restrict ourselves only to the following categories: CATEGORY, NUMERUS, PERSON, PTB_TAG and TENSE.

[4]https://dumps.wikimedia.org/

References

Duygu Altinok. 2018. Demorphy, german language morphological analyzer. *arXiv:1803.00902*.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*.

word	opposite-gender form
alla	[al, allo]
alle	[agli, ai]
colei	[colui]
costei	[costui]
esse	[essi]
dalla	[dallo, dal]
della	[dello, del]
delle	[dei, degli]
essa	[esso, egli]
impostala	[impostalo]
impostale	[impostagli]
la	[lo, gli, il]
lei	[lui]
nella	[nel, nello]
nelle	[nei, negli]
ognuna	[ognuno]
provocatele	[provocategli]
qualcuna	[qualcuno]
dalle	[dagli, dai]
sulla	[sullo, sul]
sulle	[sui, sugli]
una	[un, uno]
un’	[un]
le	[lo, i, li]
riformatorie	[riformatori]
scatenatasi	[scatenatosi]
scatenatesi	[scatenato]
tutorie	[tutorio]
al	[alla]
agli	[alle]
ciascuno	[ciascuna]
colui	[colei]
costui	[costei]
dagli	[dalle]
dallo	[dalla]
dei	[delle]
dello	[della]
essi	[esse]
esso	[essa]
i	[le]
impostagli	[impostale]
impostalo	[impostala]
li	[le]
lo	[la, le]
lui	[lei]
nei	[nelle]
nel	[nella]
ognuno	[ognuna]
provocategli	[provocatele]
qualcuno	[qualcuna]
sui	[sulle]
sullo	[sulla]
ultro	[ultra]
ai	[alle]
allo	[alla]
dai	[dalle]
dal	[dalla]
degli	[delle]
del	[della]
egli	[essa]
gli	[la]
il	[la]
negli	[nelle]
sugli	[sulle]
sul	[sulla]
nello	[nella]
un	[una, un’]
uno	[una]

Table 1: word: opposite-gender-form mapping for Italian.

word	lemma
un’, un, una, uno	lemma1
la, lo, gli, il	lemma2
le, i, li	lemma3
alla, al, allo	lemma4
alle, agli, ai	lemma5
colei, colui	lemma6
costei, costui	lemma7
esse, essi	lemma8
dalla, dallo, dal	lemma9
della, dello, del	lemma10
delle, dei, degli	lemma11
essa, esso, egli	lemma12
impostala, impostalo	lemma13
impostale, impostagli	lemma14
lei, lui	lemma15
nella, nel, nello	lemma16
nelle, nei, negli	lemma17
ognuna, ognuno	lemma18
provocatele, provocategli	lemma19
ualcuna, qualcuno	lemma20
dalle, dagli, dai	lemma21
sulla, sullo, sul	lemma22
sulle, sui, sugli	lemma23
riformatorie, riformatori	lemma24
scatenatasi, scatenatosi	lemma25
scatenatesi, scatenato	lemma26
tutorie, tutorio	lemma27
ciascuno, ciascuna	lemma28
ultro, ultra	lemma29

Table 2: word:lemma mapping for Italian.

Italian		German	
palla (ball-feminine)		diamant (diamond-masculine)	
Orig	Debias	Orig	Debias
pallina	pallina	smaragd	smaragd
biglia (ball)	pallone	topas	korund
racchetta	canestro (basket)	ultramarin	topas
pallone	rimbalzo (rebound)	salmiak	ultramarin
stecca (splint)	racchetta	*grünspan* (verdigris)	vitriol
bilia (marble)	calciato (kicked)	vitriol	saphir (sapphire)
dapples	schivata	korund	salmiak
bandierina (pennant)	guantone (mitt)	*titan* (titanium)	perle (pearl)
schivata	battitore	aquamarin	aquamarin
battitore	calciando (kicking)	*bornitrid* (boron nitride)	unedlen (base)

Table 3: Examples of top-10 nearest neighbor lists for words in Italian and in German, before and after debiasing. In red (italic) are words that were removed from the list, and in blue (underlined) are words that were added to it. Translations to English (Google Translate) for the changed words are in parenthesis, when different from source.