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

Hila Gonen¹  Yova Kementchedjhieva²  Yoav Goldberg¹,³
¹Department of Computer Science, Bar-Ilan University
²University of Copenhagen
³Allen 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 dump⁴ 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).

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.

¹http://tools.sslmit.unibo.it/doku.php?id=resources:morph-it
²https://github.com/DuyguA/DEMorphy
³Since 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: CATE-

GORY, NUMERUS, PERSON, PTB_TAG and TENSE.

⁴https://dumps.wikimedia.org/
| word         | opposite-gender form |
|--------------|----------------------|
| alla         | [al, allo]           |
| alle         | [agli, ai]           |
| colei        | [colei]              |
| costei       | [costei]             |
| esse         | [essi]               |
| dalla        | [dallo, dal]         |
| della        | [delo, del]          |
| delle        | [dei, degli]         |
| essa         | [esso, egli]         |
| impostala    | [impostalo]          |
| impostale    | [impostagl]          |
| la           | [lo, gli, il]        |
| lei          | [lui]                |
| nella        | [nel, nello]         |
| nelle        | [nei, negli]         |
| ognuna       | [ognuno]             |
| provocategli | [provocategli]       |
| qualcuno     | [qualcuno]           |
| dalle        | [dagli, dai]         |
| sulla        | [sullo, sul]         |
| sull'        | [sui, sugli]         |
| una          | [un, uno]            |
| un'          | [un]                 |
| le           | [lo, l, li]          |
| riformatì    | [riformatori]        |
| scatenatasi  | [scatenatasi]        |
| scatenatesi  | [scatenato]          |
| tutorì        | [tutorio]           |
| al           | [alla]               |
| agli         | [agli]               |
| ciascuno     | [ciascuno]           |
| colei        | [colei]              |
| costei       | [costei]             |
| dagli        | [dalle]              |
| dallo        | [dallo]              |
| dei          | [delle]              |
| dello        | [delo]               |
| essi         | [esse]               |
| esso         | [essa]               |
| i            | [le]                 |
| impostagli   | [impostalo]          |
| impostalo    | [impostala]          |
| li            | [la, le]             |
| lui          | [lei]                |
| nei          | [nelle]              |
| nel          | [nella]              |
| ognuno       | [ognuno]             |
| provocategli | [provocategli]       |
| qualcuno     | [qualcuno]           |
| sù            | [sul]                |
| sullo        | [sulla]              |
| ultro        | [ultra]              |
| ai           | [ai]                 |
| alto         | [al]                 |
| dai          | [dai]                |
| dal          | [dallo]              |
| degli        | [delle]              |
| del          | [delo]               |
| egli         | [esa]                |
| gli          | [ila]                |
| il           | [ila]                |
| negli        | [nelle]              |
| sugli        | [sul]                |
| 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, lì         | lemma3             |
| alla, al, alto    | lemma4             |
| alle, agli, ai    | lemma5             |
| colei, colui      | lemma6             |
| costei, costui    | lemma7             |
| esse, essi        | lemma8             |
| della, dallo, dal | lemma9             |
| della, delo, del  | lemma10            |
| delle, dei, degli | lemma11            |
| essa, esso, egli  | lemma12            |
| impostata, impostalo | lemma13           |
| impostale, impostagl | lemma14          |
| lei, lui          | lemma15            |
| nella, nel, nello | lemma16            |
| nelle, nei, negli | lemma17            |
| ognuna, ognuno    | lemma18            |
| provocategli, provocategli | lemma19 |
| qualcuno, qualcuno | lemma20            |
| dalle, dagli, dai | lemma21            |
| sulla, sullo, sul | lemma22            |
| surle, 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             |              |
|--------------|--------------|--------------------|--------------|
| pallina      | pallina      | smaragd            | smaragd      |
| *biglia*     | pallone      | topas              | korund       |
| racchetta    | canestro     | ultramarin         | topas        |
| pallone      | *rimbalzo*   | salmiak            | ultramarin   |
| *stecca*     | racchetta    | grünspan           | vitriol      |
| *bilia*      | calciatore   | vitriol            | saphir       |
| *dapples*    | schivata     | korund             | salmiak      |
| *bandierina* | guantone     | *titan*            | *perle*      |
| schivata     | battitore    | aquamarin          | aquamarin    |
| battitore    | *calciando*  | *bornitrid*        | *unedlen*    |

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.