**Appendix B.** Percentage of the discarded data.

All items where the coders did not agree on the status of a line (social vs non-social) or a finer-grained annotation (see Table 1) were discarded from the analysis. Below, we describe the percentage of data that was discarded.

In terms of sex in the first analysis, we discarded 44,223 (25%) observations from the dataset that females produced and 31,680 (31%) that were produced by males, as well as 307,175 (26%) word tokens from females and 236,422 (27%) from males. In the second, we discarded 66,610 (39%) observations from females and 42,308 (41%) from males, and in terms of word tokens, 528,052 (36%) of those produced by women and 347,623 (40%) by men. Concerning the level of education, we discarded between 17,943 (27%, high school) and 3,095 (36%, vocational) lines in terms of lines and between 360,469 (23%, higher education) and 8,387 (36%, children without education) of word tokens, both in the first analysis. We discarded between 25,131 lines (38%, higher education) and 74,277 (41%, higher education), as well as between 210,838 (35%, high school) and 10,215 (44%, children without education) word tokens. Finally, concerning age groups, the data set in the first analysis was reduced by between 22,582 (26%, age group 20–24) and 1,489 (35%, under 10 years of age) lines, as well as by between 164,680 (21%) and 6,889 (35%) word tokens within the same age groups. A similar pattern applied to the second analysis: we discarded the smallest percentage of observations from the age group 20‒24 (32,510 lines, 37%; 270,670 word tokens, 35%) and the most from the age group of adults over 40 years of age (3,662 lines, 42%) and children under the age of 10 (8,377 word tokens, 43%).