Chiara Zanini[1, 2], Giorgio Arcara[1], Francesca Franzon[2]

[1]Department of Neuroscience NPRSS, Università degli Studi di Padova; [2]Department of Linguistic and Literary Studies DiSLL, Università degli Studi di Padova.

Measuring the distribution of mass and count nouns. A comparison between a rating study and a corpus based analysis.

Introduction.

Despite the huge amount of literature involving theoretical linguistic and experimental approaches concerning the mass and count issue (for a review: Bale & Barner, 2011), few studies have dealt with measuring the distribution of mass and count nouns in the actual use of language (for English: Katz & Zamparelli, 2012). As a first step to fill this gap, a survey on subjective frequency of plural and singular word forms was conducted. Results of the survey were compared to frequencies collected from different corpora of written Italian.

A complementary approach.

In this study we focused on a list of 478 nouns (239 nouns inflected both at the singular and at the plural), that according to theoretical definition and to a preliminary corpus study are spaced across the range of mass and count usage. The list included even the plural of nouns for which only singular occurrences would be expected on a theoretical basis (pure mass nouns).

In the first part of the study, a questionnaire on subjective frequency was designed. Participants were asked to rate (on a 5-point Likert scale) the subjective frequency of a list of nouns. The available ratings ranged from "never heard or seen" (score: 0) to "I hear or see this word more than once a day" (score: 4). The questionnaire was administered online. A total of 126 participants took part to the study. Participants varied widely in age and education.

In the second part of the study, we collected the frequency values of the experimental stimuli from several corpora of written Italian: Colfis, La Repubblica, Subtlex (tagged), Subtlex (ignoring tags), it-WAC. These corpora differ from each other as for the strictness of the criteria concerning the texts they selected as sample.

Results on estimated subjective frequency and on corpora frequency were then compared, by means of correlations. We considered that the subjective frequency was a better estimate of the actual acceptability of the singular/plural inflection of a noun. Eventually, by investigating these correlations, it was possible to measure the reliability of corpus-based estimates on the actual possibility of occurrence of singular and plural forms of nouns.

Observed correlations with the set of experimental nouns ranged from 0.70 to 0.78. The Subtlex (ignoring tags) showed the higher correlation, whereas the Colfis showed the lowest correlation. In a further analysis we focused on nouns for which the mean subjective estimates of frequency had a score equal or greater than 2. We assumed that such threshold points to nouns for which there is a consistent judgment of plausibility across the participants of the questionnaire. Interestingly, focusing on these nouns some corpora showed zero occurrences (Colfis = 70, Subtlex = 23, Subtlex – no tag = 10).

Conclusions.

The results of the rating questionnaire showed that for almost every word in our list the plural was heard or read by the participants at least once in their life. In line with recent studies, this confirms the elasticity of mass and count usage in natural language (Katz & Zamparelli, 2012). Some corpora seem to capture better than others the distribution of possibilities that are largely accepted by the native speakers. This suggests that corpus studies on mass and count nouns should be interpreted cautiously, in the light of the corpus characteristics. Indeed, these results suggest that different selection criteria of the corpus could lead to less precise estimates of the possibility of observing a given word form in the language usage.

References.

- Bale, A., & Barner, D. (2011). *Mass-Count Distinction*. Oxford Bibliographies Online, http://ladlab.ucsd.edu./pdfs/BB@Oxford.pdf

- Baroni, M. et al. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3): 209-226

- Baroni, M. et al. (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*. http://sslmit.unibo.it/repubblica

- Bertinetto, P.M. et al. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. http://linguistica.sns.it/CoLFIS/Home.htm

- Crepaldi, D. et al. (2013). *SUBTLEX-IT: A frequency list for Italian based on movie subtitles*.

- Katz, G. & Zamparelli, R. (2012). Quantifying Count/Mass Elasticity. Choi, J. et al. (eds). *Proceedings of the 29th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, 371-379.