A method for investigating the mass-count distinction on a larger scale

In contrast to mass nouns (*blood*), count nouns (*dog*) are assumed to allow plural marking and modification by numerals (cf. Allan, 1980; for overviews cf. Katz & Zamparelli, 2012, and Pelletier, 2012). A problem that complicates a clear distinction between the two noun types is polysemy: most nouns exhibit several meaning variants which can be of different types (e.g. *glass*: a. tumbler [count], b. material [mass]). Corpus-linguistic studies of the mass-count distinction are still rare, especially those that take polysemy into account.

The first goal of this talk is to present a carefully developed step-by-step procedure based on manual semantic annotation of nouns as mass or count in German texts. For this, three components were developed:

1. A corpus of 200.000 words of contemporary German was set up, balanced along the lines of the British National Corpus. The corpus was analysed with the open-source parser MATE which identified roughly 50.000 noun tokens. These form the basis of the semantic annotation.
2. The goal of the first part of the annotation was two-fold: first, noun uses in fixed expressions were to be identified for a separate analysis. Three German native speakers were presented with a noun token in its context of utterance and each annotator individually decided whether the given noun token was part of an idiomatic expression (*Aus die Maus!* 'over and done'), an expression with a fixed meaning (*blaues Auge* 'black eye'), a proper name (*Maria*) or a complex proper name (*Berliner Mauer* 'Berlin Wall'). The second goal consisted in selecting appropriate paraphrases for the given meaning variants of the remaining noun tokens. The annotator consulted the "Duden", the standard dictionary of German, which lists paraphrases for meaning variants with examples (*Zug* for example is listed with 26 meaning variants, one of them being 'train'). The annotator then decided whether one of the paraphrases fits the given meaning and if so, the paraphrase is stored. Altogether, roughly 10.000 paraphrases were identified in the corpus.
3. The second part aimed at classifying the collected meaning variants as mass or count meanings. The annotator was presented with the paraphrase only; the original context of utterance was not given to avoid interferences between lexical meaning and interpretation in the context of utterance. Based on the paraphrase, the annotator assessed
    a. whether the combination of the meaning variant with *etwas* 'some' (*etwas Wasser* 'some water') is acceptable or not,

b. whether the combination of the meaning variant with *zwei* ‘two’ (*zwei Steine* ‘two stones’) is acceptable or not and
c. the most plausible interpretation of the expression in (b) from a set of five options.

For cases of (a) and (b), the assessment was performed on a five-step scale in order to distinguish between clear and rather blurry cases. The resulting classes allow a more fine-grained differentiation of the meanings of nouns with respect to the mass-count distinction.

The second goal of this talk is to present the results of the investigation of the German corpus to which the method has been applied.

References

Allan, K. (1980). Nouns and countability. *Language, 56*, 541–567.

Duden. http://www.duden.de

Katz, G., & Zamparelli, R. (2012). Quantifying Count/Mass Elasticity. In Choi, J., Hogue, E. A., Punske, J., Tat, D., Schertz, J., & A. Trueman (Eds.), *Proceedings of the 29th West Coast Conference on Formal Linguistics* (pp. 371-379). Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2723.

Pelletier, F. J. (2012). Lexical Nouns are Both +MASS and +COUNT, but They are Neither +MASS nor +COUNT. In Massam, D. (Ed.), *A Cross-Linguistic Exploration of the Count-Mass Distinction* (pp. 9-26). Oxford: Oxford University Press.