Determining Countability Classes

Scott Grimm, University of Rochester

Because of pioneering studies such as Jespersen (1924) and Allan (1980), it has been well-established that the data for countability is quite complex, even if it is often discussed in terms of a binary or ternary distinction. Thus, one of the outstanding, but lesser discussed, obstacles for the countability literature is how to determine exactly how many categories of countability are grammatically present and, further, how sharp the boundaries between the categories are. Recent efforts have been made to gain a broader view on countability, namely Kulkarni et al. (2013) and Kiss et al. (2016), both of which made use of human participants to rate whether a given noun satisfied a semantic or syntactic property. This paper presents a complimentary approach which explores inducing countability classifications directly from corpus data, which can then be compared (where possible) with the results of those researchers.

Database This study is based on a large portion ($\sim$ 150 million words) of the Corpus of Contemporary American English (COCA) corpus (Davies, 2008), which was converted into a set of data frames, one for each noun, subsequent to applying various NLP tools (sentence tokenization, dependency parsing). Each noun data frame underwent semi-automatic annotation for a range of variables of interest (at present 92 different variables). These variables include straightforward grammatical information about a noun's immediate environment, such as a given noun's rate of pluralization, if it appears with a particular quantifier (e.g. *such*), and what that rate of occurrence is. In addition, higher-level information of a noun's combinatorics is tracked, such as what types of adjectives or verbs it co-occurs with, for which the "supersenses" of Wordnet (Fellbaum, 1998) are taken as an adequate representation. Taken together, this depth of data and range of variables permits understanding variation in countability from a new perspective. I discuss several results from studies on a preliminary "countability database", which aggregates the information for a sample of 2000 nouns from the data.

Predicting (Non-)Count Status The data frames of different noun's co-occurrence patterns obviously contain many correlated variables, for instance, a noun which robustly occurs with the quantifier *many* might be expected to also occur with the quantifier *several*. To assess which grammatical characteristics might prove most predictive of (non-)count status, a series of regression models were run, as well as random forest classification, to assess the independent contribution of the different variables. This is of some theoretical interest: Chierchia (2010) very plausibly discusses impossibility of combining with numerals as the 'signature property' of non-countable nouns. To examine this question, the countability database was extended with the countability classification of nouns provided in the CELEX database (Baayen et al., 1996), which for present purposes is taken as a gold standard. CELEX's countability classification then is used as the dependent variable. The models reveal that the most reliable trait of countable nouns is occurring as a bare plural and likewise the most reliable trait of non-countable nouns is occurrence as bare singulars. The other significant predictors are the quantifier *several*, a predictor of countable nouns, and *some* and *such*, predictors of non-countable nouns, while other grammatical traits, including co-occurrence with cardinals are not significant. These results provide evidence for examining grammatical properties of nouns independently: i.e., indefinite determiners may relate to countability classification in a way close to, yet distinct from that of plural inflections.

Inducing Countability Classes Various machine learning techniques (k-means clustering, Bayesian hierarchical clustering, partitioning around medoids, etc.) on the countability database were applied so as to establish countability classes through purely grammatical characteristics. Using the measurements from the most predictive variables and analyzing the k-means clustering results indicate that 4 clusters are optimal; however, exploratory examinations of larger values for k turn out to be informative. For instance, Allan (1980) considered *cattle* as a representative of a countability class. When examining clusters with k=9 or higher, it turns out that *cattle* is the sole member of a cluster, i.e., *cattle* is an extreme outlier in terms of its grammatical behavior. In the talk, I will illustrate the "degrees of countability" achieved through different clusterings, both where they illuminate issues and where they obscure them.

Ongoing Work Current progress is being made in two directions. First, sense disambiguation methods are being employed to convert the database from a set of words to a set of word senses. Second, the database is being combined with psycholinguistic databases which provide metrics of the concrete/abstract distinction (Brysbaert et al., 2014; Troche et al., 2014) to address the interaction between the count/non-count and concrete/abstract distinction.

References

Keith Allan. Nouns and countability. *Language*, 56(3):41–67, 1980.

R.H. Baayen, R. Piepenbrock, and L. Gulikers. *CELEX2*. Linguistic Data Consortium, Philadelphia, PA, 1996.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014.

Gennaro Chierchia. Language, thought, and reality after Chomsky. In Julie Franck and Jean Bricmont, editors, *The Chomsky Notebook*, pages 142–169. Columbia University Press, New York, 2010.

Mark Davies. The Corpus of Contemporary American English: 450 million words, 1990-present, 2008. Available online at http://corpus.byu.edu/coca/.

Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, 1998.

Otto Jespersen. *The philosophy of gammar*. Allen & Unwin, London, 1924.

Tibor Kiss, Francis Jeffry Pelletier, Halima Husic, Roman Nino Simunic, and Johanna Marie Poppek. A sense-based lexicon of count and mass expressions: The Bochum English countability lexicon. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 2016. European Language Resources Association (ELRA).

R. Kulkarni, S. Rothstein, and A. Treves. A statistical investigation into the cross-linguistic distribution of mass and count nouns: Morphosyntactic and semantic perspectives. *Biolinguistics*, 7:132–168, 2013.

Joshua Troche, Sebastian Crutch, and Jamie Reilly. Clustering, hierarchical organization, and the topography of abstract and concrete nouns. *Frontiers in Psychology*, 5:360, 2014.