
Graham Katz, Roberto Zamparelli
Georgetown University, Università degli Studi di Trento
egk7@georgetown.edu, roberto.zamparelli@unitn.it

Meaning-shifting plurality and the Count/Mass Distinction

Introduction

The semantics of plurality is a recurring theme in the formal linguistics literature (Link 1983; Krifka 1989; Schwarzschild 1996; Landman 2000; Heycock & Zamparelli 2005). With the exception of a few frequently-cited idiosyncratic cases (*brain/brains*, *glass/glasses*, see (Acquaviva 2008)) it is assumed that the meaning of a plural noun is closely related to that of the corresponding singular. The goal of this talk is to test this assumption using the quantitative tools of vector space (or “distributional”) semantic models (DSM) (Landauer & Dumais 1997; Baroni & Lenci 2010). When applied to the singular and plural forms of the same word, these methods clearly show that there is a large proportion of nouns whose distribution in the singular and the plural differs, in terms of neighboring content words. In these nouns, plurality seems to come with a meaning shift that goes beyond number, and which is detected by the DS methods.

Further analyses show that the shift in the interpretation of plurals correlates with the *countability* of the noun. Nouns with closely related plural and singular forms tend to be count nouns (nouns that occur in canonical *count* contexts (Baldwin & Bond 2003)), nouns with meaning-shifted plurals tend to be mass nouns. The experiment thus offers a novel angle for examining the notoriously elusive count-mass distinction (Quine 1960; Pelletier & Schubert 1989; Chierchia 2010), and a clear example of what many DSMs miss by studying word distributions in terms of their *lemma* alone.

Singular/Plural Distance

To study noun-number semantic differences, we built a vector space model of the 100 million token UKWAK-1 corpus (Baroni et al. 2009) for the 20,000 most frequent content words in the corpus, using the COALS algorithm (Rohde et al. 2005). Our model maintained both part of speech and lemma information. We analyzed the 2131 noun-types which appeared in both singular and plural forms in the model.

We chose two ways to examine semantic proximity. First using a vector cosine measure and then using a word-based measure. The 25 nouns with lowest and highest singular/plural cosine-similarity are listed below:

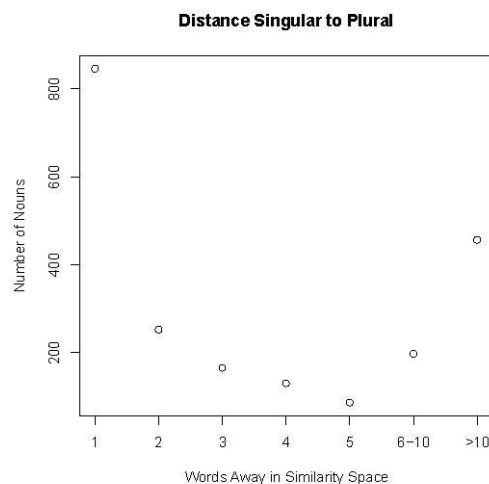
Low plural/singular similarity: *leave make creator con humanity extreme good disadvantage toddler strength fortune horizon total story hip mouse dozen tip monkey security term medium support manner custom*

High plural/singular similarity: *phone therapist resort impairment reaction list century speaker cookie engine locomotive pool sector cancer bomber venture guitar beach int examiner thou squirrel tale surgeon acid*

As we see, the nouns with lowest sing/plural similarity contain ambiguous nouns where different meanings are associated with one of the two forms (e.g. *mouse/mice*, *good/goods*, *security/securities*).

To investigate this contrast, we used the cosine-similarity metric to determine how many words in the model were distributionally closer to the singular than the corresponding plural was. In our model, *boxes* was the nearest neighbor of *box*, but *flight* comes between *airline* and *airlines*, and *arms* was more than 10 words away from *arm*. In Figure 1 we display the distribution of nouns by distance in words:

FIGURE 1. WORD DISTANCE.



For more than a third of the nouns under study the closest element in the model to the singular form is, as expected, the corresponding plural form (we call these the “near” group). Interestingly, however, for more than a fifth of the nouns the plural isn’t among the 10 closest neighbors of the singular (we call these the “far” group). The average sing/plural cosine similarity of “far” group is 0.18, while the average similarity for “near” group is 0.50.

Mass/Count

In the formal literature, it is often noted (Carlson 1977; Chierchia 1998; Rothstein 2010) that mass nouns such as *wine* undergo a semantic shift when they are used in count contexts or pluralized (*two wines* means two kinds of wine). This suggests the hypothesis that the “far” nouns would be more count nouns and that the near nouns would be more mass nouns. To investigate this, we examined the rate of occurrence of the nouns in the sample in typical mass contexts (e.g. with *much*) and in typical count contexts (e.g. with *every*).

Looking first at mass contexts we find that the average rate of mass contexts occurrence (by noun type) for the 'far' nouns is significantly higher than that for 'near' nouns (0.0028 vs. 0.0014, $p = 0.003041$). Even more dramatic is the difference in the distribution of count-context rates between the two groups: The difference between group means (0.192 vs. 0.149) is highly significant ($p = 7.09e-09$). In short, the 'near' group contains more predominantly count nouns than the 'far' group.

FIGURE 2. RATE OF COUNT CONTEXTS FOR 'NEAR' NOUNS.

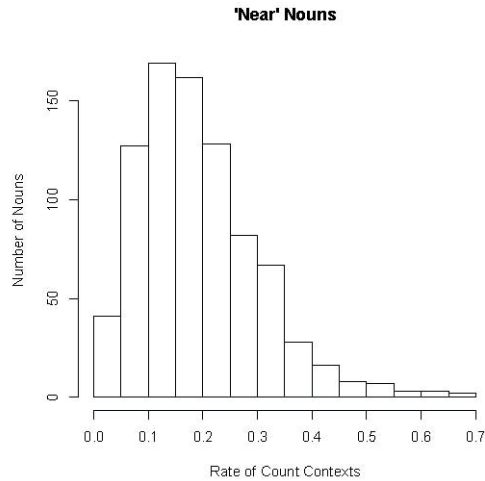
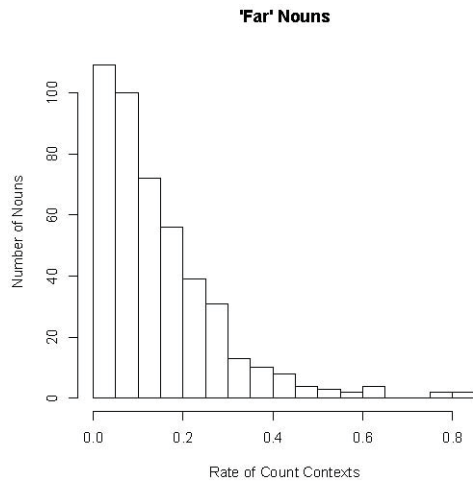


FIGURE 3. RATE OF COUNT CONTEXTS FOR 'FAR' NOUNS.



Conclusions

It is clear, then, that in building vector-space semantic models, some word-form information should be maintained. Collapsing lemmas in vector-space models ignores important semantic information. This is particularly the case for the class of mass nouns.

References

- Acquaviva, P. 2008. *Lexical plurals: a morphosemantic approach*. Oxford University Press, USA.
- Baldwin, T. and Bond, F. 2003. Learning the countability of English nouns from corpus data. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 463-470, Morristown, NJ, USA. Association for Computational Linguistics.
- Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209-231.
- Baroni, M. and Lenci, A. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*.
- Carlson, G. N. 1977. *Reference to Kinds in English*. PhD thesis, University of Massachusetts, Amherst, Amherst, MA.
- Chierchia, G. 1998. Plurality of mass nouns and the notion of semantic parameter. In *Events and grammar*, Rothstein, S. (ed.). Kluwer Academic, Dordrecht.
- Chierchia, G. 2010. Mass nouns, vagueness and semantic variation. *Synthese* 174(1), 99-149.
- Heycock, C. and Zamparelli, R. 2005. Friends and colleagues: Plurality, coordination, and the structure of DP. *Natural language semantics* 13(3), 201-270.
- Krifka, M. 1989. Nominal reference, temporal constitution and quantification in event semantics. In *Semantics and Contextual Expression*, Bartsch, R., van Benthem, J., and van Emde Boas, P. (eds.). Foris Publications, Dordrecht.
- Landauer, T.K. and Dumais, S.T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2), 211-240.
- Landman, F. 2000. *Events and plurality: The Jerusalem lectures*. Kluwer Academic Publishers.
- Link, G. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. *Meaning, Use, and Interpretation of Language*, 302-323. de Gruyter, Berlin.
- Pelletier, F. and Schubert, L. 1989. Mass expressions. *Handbook of philosophical logic*, 4:327-407.
- Quine, W.V.O. 1960. *Word and Object*. The MIT Press, Cambridge, MA.
- Rohde, D., Gonnerman, L. and Plaut, D. 2005. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. Unpublished Manuscript. Available online at <http://tedlab.mit.edu/~dr/Papers/RohdeGonnermanPlaut-COALS.pdf> [March 05, 2011]
- Rothstein, S. 2010. Counting and the Mass/Count Distinction. *Journal of Semantics* 27(3), 343.
- Schwarzschild, R. 1996. *Pluralities*. Dordrecht: Kluwer.
-