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 then 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.**

![Near Nouns Graph]

**Figure 3. Rate of Count Contexts for ‘Far’ Nouns.**

![Far Nouns Graph]

**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


