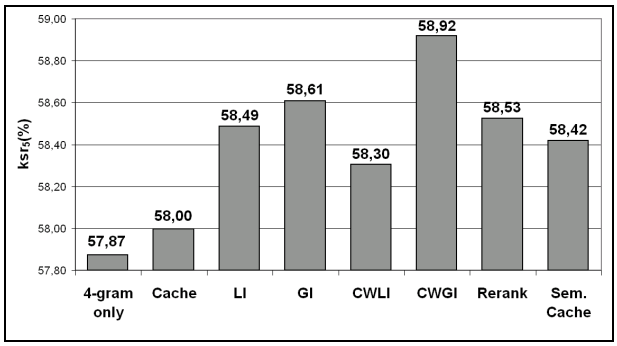

Figure 3: Results (ksr_5) for all methods tested.

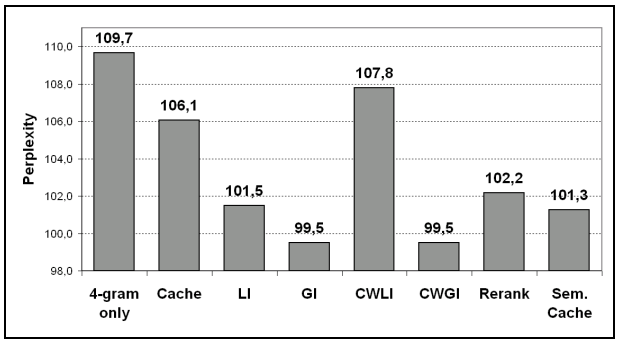

Figure 4: Results (perplexity) for all methods tested.

Using the results of our 8 samples, we performed paired *t* tests for every method with the baseline as well as with the cache model. All gains for *ksr* turned out to be highly significant (sig. level < 0.001), and apart from the results for CWLI, all perplexity reductions were significant as well (sig. level < 0.007), with respect to the cache results. We can therefore conclude that, with exception of CWLI, all methods tested have a beneficial effect, even when compared to a simple cache model. The highest gain in *ksr* (with respect to the baseline) was obtained for the confidence-weighted geometric interpolation method (CWGI; +1.05%), the highest perplexity reduction was measured for GI as well as for CWGI (-9.3% for both). All other methods (apart from IWLI) gave rather similar results (+0.6 to +0.8% in *ksr,* and -6.8% to -7.7% in perplexity).

We also calculated for all samples the correlation between *ksr* and perplexity. We measured a *Pearson* coefficient of -0.683 (Sig. level < 0.0001).

At first glance, these results may not seem overwhelming, but we have to take into account that our *ksr* baseline of 57.9% is already rather high, and at such a level, additional gains become hard to achieve (cf. Lesher et al, 2002).

The fact that CWLI performed worse than even simple LI was not expected, but it can be explained by an inherent property of linear interpolation: If one of the models to be interpolated overestimates the probability for a word, the other cannot compensate for it (even if it gives correct estimates), and the resulting probability will be too high. In our case, this happens when a word receives a high confidence value; its probability will then be overestimated by the LSA component.

5 Conclusion and further work

Adapting a statistical language model with semantic information, stemming from a distributional analysis like LSA, has shown to be a non-trivial problem. Considering the task of word prediction in an AAC system, we tested different methods to integrate an n-gram LM with LSA: A semantic cache model, a partial reranking approach, and some variants of interpolation.

We evaluated the methods using two different measures, the keystroke saving rate (*ksr*) and perplexity, and we found significant gains for all methods incorporating LSA information, compared to the baseline. In terms of *ksr* the most successful method was confidence-weighted geometric interpolation (CWGI; +1.05% in *ksr*); for perplexity, the greatest reduction was obtained for standard as well as for confidence-weighted geometric interpolation (-9.3% for both). Partial reranking and the semantic cache gave very similar results, despite their rather different underlying approach.

We could not provide here a comparison with other models that make use of distributional information, like the trigger approach by Rosenfeld (1996), Matiasek and Baroni (2003) or the model presented by Li and Hirst (2005), based on *Pointwise Mutual Information* (PMI). A comparison of these similarities with LSA remains to be done.

Finally, an AAC system has not only the function of simple text entering but also of providing cognitive support to its user, whose communicative abilities might be totally depending on it. Therefore, she or he might feel a strong improvement of the system, if it can provide semantically plausible predictions, even though the actual gain in *ksr* might be modest or even slightly decreasing. For this reason we will perform an extended qualitative