Building a Lexical Database for an Interactive Joke-Generator
Ruli Manurung, Dave O’Mara, Helen Pain, Graeme Ritchie, Annalu Waller, Rolf Black

Data Preparation
Database: all lexical resources stored in a relational database.

WordNet+Unisyn: disambiguation using POS tag, handling of compound nouns.

Phonetic relations: phonetic similarity (edit distance-based), rhymes, spoonerisms precomputed.

Manual disambiguation: Widgit conceptcodes for picture symbols attached to about 10k word senses, by hand. The common senses of wordforms which appeared in Schonell spelling list were selected by hand.

Auto disambiguation: disambiguated Widgit & Schonell data used to select senses for wordforms in other sources (e.g. MRC database).

Familiarity scoring: prioritized sources (MRC>Schonell>Widgit>Semcor) combined to rate the familiarity of each word-sense: an “F-score”.

Technical details
The lexical knowledge base:
• is implemented as a PostgreSQL database
• is accessible via a Java API
• contains symbolic links to the pictures
• should be available from November 2006 (see STANDUP website)
• is currently used in STANDUP system for language-impaired children:

References