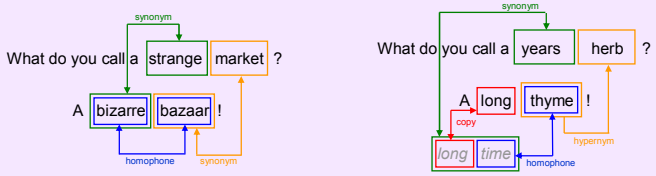
Building a Lexical Database for an Interactive Joke-Generator

Ruli Manurung, Dave O'Mara, Helen Pain, Graeme Ritchie, Annalu Waller, Rolf Black

Functional requirements

We are building software which allows language-impaired children to interactively construct their own jokes (simple punning riddles). An underlying joke generation system is able to automatically generate jokes such as:



This generator requires a lexicon with the following information:

- Part-of-speech (POS) tags
- Phonetic spelling, for computing:
  - homophones time ↔ thyme
  - rhyme pub ↔ tub
  - spoonerism bare/spank ↔ spare/bank
- Compound nouns and their components, e.g. long time, red herring, traffic jam
- Distinct senses of a word/phrase, e.g. match=sporting event, match=ignition stick
- Semantic relations:
  - synonyms strange ↔ bizarre
  - hypernyms thyme → herb
  - meronyms car → wheel

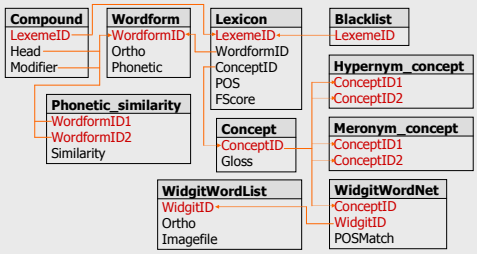
Resources used

- **WordNet**: >200k word senses, synonyms (**synsets**), hypernym hierarchy, meronyms.
- **Unisyn**: pronunciation dictionary, assigning phonetic strings to >115k word forms. Edinburgh accent used.
- **Widgit conceptcodes**: >11k concepts linked to >6k Widgit Rebus symbols, >4k Mayer-Johnson PCS symbols.
- **SemCor**: subset of Brown corpus with >230k WordNet sense-tagged words. >35k WordNet entries have SemCor frequency>0.
- **Schonell spelling lists**: spelling list of >3k words for children aged 7-12. Used as preferred source of "familiar" words.
- **MRC Psycholinguistic Database**: various ratings relevant to familiarity.
- **BNC Spoken Corpus**: frequency ratings for compound nouns.

The Lexical Knowledge Base

Core lexicon:

- 130k lexemes
  - ✓Sense (synset)
  - ✓Part of speech
  - ✓F-score (in [0,1])
- 79k wordforms
  - ✓Orthography
  - ✓Phonetic spelling
- 32k compound nouns
  - ✓Head
  - ✓Modifier
- 85k concepts
  - ✓WordNet gloss
- 65k hypernym pairs, 7.5k meronym pairs
- 10k Widgit-to-WordNet matches (>8k POS matched)
- >500k phonetic similarity ratings (in [0.75,1])



Example:

Lexicon

LexemeID	WordformID	ConceptID	POS	FScore
lx141161	wf117922	cn114381813	n	0.105
lx141159	wf150640	cn114381813	n	0.786
lx071108	wf103830	cn107348541	n	0.208
lx071018	wf046732	cn107338521	n	0.723

Wordform

WordformID	Ortho	Phonetic
wf117922	years	y * ir; r z
wf046732	herb	h * er r b
wf103830	thyme	t * ai m
wf104019	time	t * ai m
wf058637	long	l * oo ng

Hypernym_concept

ConceptID1	ConceptID2
cn107338521	cn107348541
"herb"	"thyme"

Compound

LexemeID	Head	Modifier
lx141159	wf104019	wf058637
"long time"	"time"	"long"

Phonetic_similarity

WordformID1	WordformID2	Similarity
wf103830	wf104019	1.0
"thyme"	"time"	

Lexical relations stored in additional cache tables:

- **Syntactic**: noun, verb, adj, mod, compound
- **Semantic**: synonym, hypernym, meronymy, alternate meaning
- **Phonetic**: homophone, rhyme, spoonerism, prefix, suffix

Supported by EPSRC *grants GR/S15402/01 and GR/R83217/01*

http://groups.inf.ed.ac.uk/standup

User requirements

Potential users and suitable experts (teachers and therapists), suggested that:

- Speech output should be available.
- Words should have a symbol attached to it, preferably from a standard AAC symbol-library, e.g.:

- Words should be grouped into subject-areas (topics), clustered into a hierarchy, for easy user access.
- Words deemed unsuitable for the target users (e.g. swear words, sexual terminology) should be avoided.
- The joke generator should prefer words likely to be familiar to users.

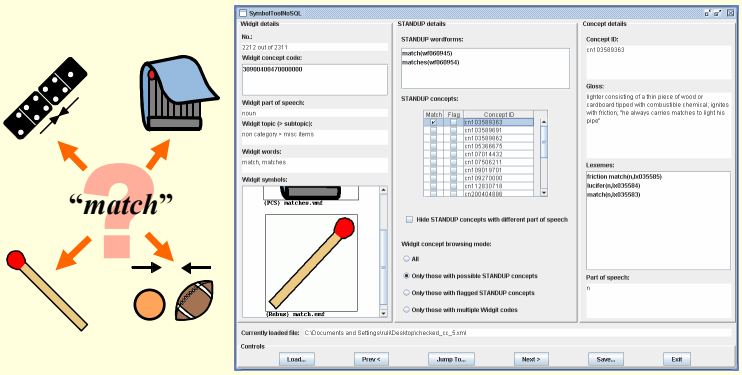
Data Preparation

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.



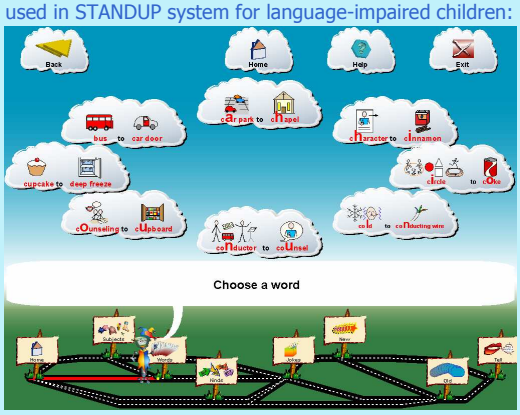
Auto disambiguation: disambiguated **Widgit** & **Schonell** data used to select senses for wordforms in other sources (e.g. **MRC** database).

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

Manurung, R., O'Mara, D., Pain, H., Ritchie, G., & Waller, A. (2005). Facilitating User Feedback in the Design of a Novel Joke Generation System for People with Severe Communication Impairment. In *HCI 2005* (CD), Vol.5, G. Salvendy (Ed). Lawrence Erlbaum, NJ, USA.

O'Mara, D., Waller, A., Ritchie, G., Pain H., & Manurung, H.M. (2004). The role of assisted communicators as domain experts in early software design. In *Proceedings of the 11th Biennial Conference of the International Society for Augmentative and Alternative Communication* (CD) Natal, Brazil, 6-10 October 2004.

Ritchie, G., Manurung, R., Pain, H., Waller, A., O'Mara,D. (2006) The STANDUP Interactive Riddle Builder. *IEEE Intelligent Systems* 21 (2), March/April. Pp. 67-69.