

Terminology, translation, and PRESEMT; word frequency lists and KELLY

Adam Kilgarriff
Lexical Computing Ltd

PRESEMT

- **EU FP7 project**
 - **FP7-ICT-4-248307**
 - **2010-1012**
- **Pattern Recognition based Statistically Enhanced MT**
- Six partners, five countries
- Languages: Czech English German Greek Italian
- Comparable Corpora BootCat (CCBC)
 - Demo by Jan Pomikalek
- <http://www.presemt.eu>

KELLY

- “Keywords for Language Learning for Young and adults alike”
- EU lifelong learning project:
 - Goal: wordcards
 - Word in one lg on one side, other on other
 - *Language learning*
 - 9 languages, 36 pairs
 - Arabic Chinese English Greek Italian Norwegian Polish Russian Sweden
 - Partners in 6 countries
- <http://su.avedas.com/converis/contract/321>



Method

- Prepare monolingual lists
- Translate
 - Each into 8 target languages
 - Professional translation services
- Integrate, finalise
- Produce cards
- Goal for each set
 - 9000 pairs at 6 levels

Stages

- Sort out corpora, tagging
- Automatically generate M1 lists
 - names, numbers, countries ...
 - keywords vis-a-vis other corpora
- Review, compare, prepare M2 lists
- Translate
- Use translations: M3 lists
- Finalise

review - how?

- points system
 - 2 points for each of 6 levels
 - 12 points for most freq words
- deduct points for words in over-represented areas
- add in words from other corpora

Translation database

- On the web
- All translations entered into it
- Queries like
 - *All Swedish words used as translations more than six times*
 - *All 1:1:1:1... 'simple cases'*

Using the translations database

- Find words not in M2 lists, that need adding
 - Multiwords
 - English *look for*
 - Probably, the translation of a high-freq word in several of the 8 other lgs
 - So:
 - ***add it to English list***
 - Homonyms: could be similar

Monolingual master lists (M3)

- Based on a WAC corpus
- Input from other same-lg corpora
- And from translations from 8 lgs
 - **Useful** words which might not be hi-freq
 - added words/multiwords must be above a lower freq threshold
- Target 9000

Matches across 9 languages

- Set of symmetrical relations across all 36 pairs
 - music
 - library
 - sun
 - hospital
 - theory

Big problems

- Multiwords (as anticipated)
- Homonymy (as anticipated)
- *orange banana alphabet elbow, **Hello***
 - Worse than anticipated
 - Lists from spoken corpora, learner corpora, needed
 - Relation between
 - Competence for communicating
 - The corpora at our disposal

(Monolingual) Word Lists

- Define a syllabus
- Which words get used in
 - Learning-to-read books (NS children)
 - NNS language learner textbooks
 - Dictionaries
 - Language testing
 - NS: educational psychologists
 - NNS: proficiency levels

Should be corpus-based

- Most aren't
 - Corpora are quite new
- Easy to do better
- People will use them
 - Maybe also Governments

How

- Take your corpus
- Count
- Voila

Complications

- What is a word
 - Words and lemmas
 - Grammatical classes
 - Numbers, names...
 - Multiwords
 - Homonymy
-
- All are slightly different issues for each lg

What is a word; delimiters

- Found between spaces
 - Not for Chinese: *segmentation*
 - English
 - *co-operate, widely-held, farmer's, can't*
 - Norwegian, Swedish
 - Compounding, separable verbs
 - Arabic, Italian
 - Clitics, *al, ...*
 - ...
-

Words and lemmas

- Word form (in text)
 - *invading*
- Lemma (dictionary headword)
 - *Invade* for forms *invade invades invaded invading*
- Lemmatisation
 - Chinese, none; English, simple
 - Middling: Swe Nor It Gr
 - Tough: Rus, Pol, Ara

Word Families

- Derivational morphology
 - *efficient/efficiently*
 - *access/accessible/accessibility*
 - *available/availability/unavailable*
- 'Word families' tradition
 - eg: Coxhead, Academic word list
 - Pedagogy: one item to learn
 - ***But***
 - Where do families end? Different meanings

Grammatical classes

- *brush* (verb) and *brush* (noun)
 - Same item or different?
 - (both in same word family)
- Required
 - (short) list of word classes
 - POS-tagger
 - Will make mistakes

Marginal cases

- Numbers
 - *twelve, seventeenth, fifties*
 - Closed sets
 - Days of week, months
 - Countries
 - Capitals, nationalities, currencies, adjectives, languages
 - regional/dialects, political groups, religions
 - *easter, christmas, islam, republican*
 - policies always needed
-

Multiwords

- *According to*
 - Linguistically a word but
 - Multiword frequency list: top item *of the*
 - Can't use freqs (alone) to select multiwords

Homonymy

- *bank* (river) and *bank* (money)
- Word sense disambiguation
 - We can't do (with decent accuracy)
 - We can't give freqs for senses
- Lists of words not meanings
 - Sometimes disconcerting

Corpora

- A fairly arbitrary sample of a lg
- To limit arbitrariness of wordlist
 - Make it ***big*** and ***diverse***
- **WACKY** corpora
 - From web
 - Can do for any language
 - ??? Comparable ???
 - Web language: less formal

Word lists are useful, but

- ...are they scientific?
 - A tiny bit, occasionally
- ...could they be scientific?
 - **Yes**
 - article of faith
 - By the end of KELLY, we'll have a clearer idea how