THE MULTIWORD EXPRESSION PROJECT

ORGANISATIONAL BACKGROUND

- 3 year project initiated in April 2001
- Funded by NTT (Japan) and YY Technologies (U.S.)
- International project involving CSLI, Cambridge University (U.K.) and NTT (Japan)
- Combines computational and linguistic interests

BASIC AIMS OF PROJECT

- To accumulate knowledge on the workings of multiword expressions, focusing on English but with an eye to cross-lingual analysis
- To map out the space of multiword expressions
- To develop a formalism for different multiword expression types, and implement those formalisms within the ERG/LKB
- To develop techniques to automatically extract multiword expressions and feed them into a grammar for parsing/generation purposes

WORKING DEFINITION

Multiword expression (MWE): phrase that is not *entirely* predictable on the basis of standard grammar rules and lexical entries

MULTIWORD EXPRESSION TYPES (1/2)

- Lexicalised phrases:
  - "words with spaces" (e.g. *ad hoc*)
  - idioms (e.g. *let the cat out of the bag*)
  - idiomatic constructions (e.g. *the Xer the Yer*)
  - verb-particle constructions (e.g. *give up*)
  - light verb constructions (e.g. *take a walk*)
  - (semantically idiosyncratic) compound nouns (e.g. *grass roots*, *tea towel*)
  - (semantically idiosyncratic) adjective–noun constructions (e.g. *little black book*)

MULTIWORD EXPRESSION TYPES (2/2)

- Institutionalised phrases (syntactically/semantically idiosyncratic):
  - compound nouns (e.g. *post office*)
  - adjective–noun constructions (e.g. *heavy drinker*)
  - verb–object pairs (e.g. *kindle excitement*)
  - ordered noun sequences (e.g. *knife and fork*)
- *Institutionalised phrases syntactically/semantically compositional but marked statistically*

IMMEDIATE PROBLEMS

- Map out different MWE types (work out what axes are required to do this – syntactic and semantic description)
- Dividing line between lexicalised and institutionalised phrases fuzzy at best (consider *fine weather*)
- Representation issue: list lexicalised phrases in lexicon, but what about institutionalised phrases?
- While it makes sense to subclassify MWEs according to construction type, what similarities and differences are there between the different subclassifications?
- Should we block productive expressions not expressly listed in the lexicon in the case that a MWE exists?
- Parsing vs. generation

PROPERTIES OF MWEs

- Semantic compositionality ("idiomatic expressions" vs. "idiomatically combining expressions") — *kick the bucket* vs. *let the cat out of the bag*
- Identifiability (encoding vs. decoding distinction: Fillmore *et al.* 1988) — *wide awake* vs. *kick the bucket*
- Grammaticality vs. syntactic irregularity — *kick the bucket* vs. *all of a sudden*
- Substantive vs. formal MWEs — *pull a swifty* vs. *either ... or ...*
- Syntactic variability

SYNTACTIC VARIABILITY

- Cline of syntactic variability from 100% syntactically frozen strings (*in particular*) to compositional phrases (*strong coffee*)
- Forms of variation
  - modifiability/quantifiability
  - topicalisability/ellidability of certain parts of the expression
  - pronominalisability of certain parts of the expression
  - passivisability (verbal expressions)
  - adverb insertability (verbal expressions)
  - coordination
  - relativisability (adjectival/verbal expressions)
  - attributive/predicative alternation (adjective–noun constructions)
  - part-wise semantic correlation with other MWEs

ISSUES WITH SYNTACTIC VARIABILITY

- How can we predict/represent the variability of a given MWE?
  - generally chronic data sparseness associated with corpus-based approaches
  - feature-based inference process seems cognitively plausible, but what features are axiomatic and how do they determine/predict non-axiomatic features?
- Better to model syntactic variability by lexical type or predict extra-lexically?

PRODUCTIVITY

- Differences between MWEs and simplex words
- Different MWE types and subtypes are associated with different levels of productivity:

  E.g. compound nouns — `made-of` basically fully productive (e.g. *cloud car*), `has-part` less so (e.g. *4-door car* vs. **sunroof car*), and instances such as *pickpocket* are non-productive

- How can we predict/constrain productivity?

COMPUTATIONAL CONSIDERATIONS

- How to represent MWEs efficiently in the ERG?
- How to construct a database that can be used with the ERG and other systems?
- How to extract MWEs?
- How to classify MWEs once extracted, to be able to feed them directly into the ERG?

OTHER (LONGER-TERM) ISSUES

- Cross-lingual analysis of MWE types/use/behaviour
- Stochastic handling of MWEs
- Evaluation (coverage, analysis, generation)

LOOKING TO THE MORE IMMEDIATE FUTURE

- Verb-particle constructions:
  - upgrade handling within ERG
  - settle upon set of lexical types covering all verb-particle instances
  - start extracting verb-particle constructions to be fed into the ERG
  - interface work with move over to a lexical database
- *Set up weekly reading group to get the ball rolling*