FAILINGS OF EXISTING COLLOCATION TECHNIQUES

- Pure statistical methods:
  - no (additional) linguistic insight into what is being extracted
  - partial coverage of different methods
  - tend to focus on N-grams (esp. 2-grams)
- Taxonomy-based substitution methods:
  - need to commit to a given taxonomy (e.g. WordNet)
  - need for WSD
  - brittle over narrow domains
- No analysis of syntactic or semantic idiosyncrasy of "collocations" (context-free analysis)
- Incompatible with syntactically-flexible collocations

"COLLOCATION" EXTRACTION FROM FIRST PRINCIPLES

- "Collocations" are a mix of lexicalised (semantically/syntactically idiosyncratic) and institutionalised (statistically marked) expressions
- Use linguistic properties of each to differentiate between the two as part of extraction
- Lexicalised expressions:
  - leave analysis of syntactic idiosyncrasy to grammar
  - employ direct analysis of semantic idiosyncrasy by looking at patterns of lexical combination/subcategorisation
- Institutionalised expressions:
  - syntactically and semantically compositional expressions which stands in opposition to anti-collocations

BASIC IDEA

1. Use parse dependencies to model contextual preferences
2. Determine synonymy by way of similarity of contextual preference
3. Use synonym sets to test substitutability of all lexical items in MWE candidate
4. Compare lexical context of MWE candidates with that of simplex headword

		Substitution alternants exist?	
		No	**Yes**
Disparate lex. context to simplex headword?	**Yes**	Lex'lised exp.	???
	No	Inst'lised exp.	Compositional

ALGORITHM

1. Chunk-parse text (Link, Apple Pie, etc.), extract dependency pairs (e.g. subject–verb, adjective–noun, verb–object, preposition–noun, ...)

2. Model the lexical context of each word type by way of dependency information

3. Model synonymy by clustering over lexical context

4. For a given MWE candidate:

   (a) search for instances of internal modification and exit if found to occur with same lexical context in sufficient numbers

   (b) generate substitution alternants by substituting in synonyms for each component word

(c) compare the relative frequency (and lexical context) of substitution alternants

(d) compare the relative frequency (and lexical context) of the simplex headword with the MWE candidate

(e) classify MWE candidate according to judgements on substitutability and similarity to simplex lexical context

WORD SENSE AND SIMILARITY OF LEXICAL CONTEXT

- Simplex headwords generally occur more often and in a wider range of lexical contexts than MWEs (partly as a result of having a broader range of senses – cf. *baggage* vs. *emotional baggage*)
- SOLUTION: compare spread of lexical context within those contexts observed for the MWE candidate

ADVANTAGES OF PROPOSED METHOD

- Domain-tunable
- Purely data-driven (independent of external resources)
- No need for WSD
- Linguistically motivated
- Compatible with syntactically fixed collocations

PRELIMINARY EXTRACTION TASK: VERB-PARTICLES

- Focus on verb-particle constructions as first extraction test case
- Relatively high frequency of occurrence, relatively well defined/underst linguistically, meaning that:
  - they are *relatively* easy to extract
  - "gold standard" repositories are available for evaluation purposes (Alvey tools, LinGO grammar(??))

VERB-PARTICLE CONSTRUCTIONS: BASELINE

- Some tag sets (e.g. Penn Treebank) distinguish between prepositions (IN/TO) and particles (RP) to some degree

  $\Rightarrow$ run (Brill) tagger over sentence-delimited, case-normalised, stemmed data and extract verb-particle constructions by way of regular expression over tag sequences
  (e.g. `V* (^ W*|COMP|.)* RP`)

- Possibility of distinguishing between `V NP Part` and `V Part NP` instances, based on analysis of what comes between the verb and the particle

- PROBLEMS/LIMITATIONS

  – high precision/low recall (only 846 verb-particle types in WSJ/Brow component of Penn Treebank: 572 of these not in combined Alvey tools/LinGO listing of verb particles)

VERB-PARTICLE CONSTRUCTIONS: DEPENDENCY BASED

- **Headword:** verb, **"modifier":** particle
- Substitution a valid test?
- Lexical context: subject–verb, verb–object (incl. the lack thereof, i.e. intransitive usages)
- If particle-final realisation possible (e.g. *phone him up*), then treat as verb-particle construction irrespective of similarity of lexical context (compositional vs. non-compositional usages)

VERB-PARTICLE CONSTRUCTIONS: DISTRIBUTION BASED

- Look first at fronted prepositional phrases (i.e. non-particle usages of prepositions) to evaluate what patterns of occurrence they display (optionally in combination with the main verb they occur with)
- Next look at NPs sandwiched between verbs and "floating" prepositions (i.e. particle-final verb-particle constructions)
- Out of all NPs which immediately follow a verb and preposition, determine which of the two groups above they are most similar to