Multiword expressions

1 Introduction

Even the best existing formal grammars of natural languages generate a large proportion of utterances that sound stilted, ugly or simply wrong to native speakers. Many of the problems are due to properties of multiword expressions: that is, phrases that are not entirely predictable on the basis of standard grammar rules and lexical entries. Most multiword expressions are not completely fixed strings but neither are they as variable as compositional phrases. There are complex interrelationships between various multiword expressions, so they cannot be satisfactorily treated in isolation of each other. Consideration of idioms leads into problems of deciding where the boundary lies between literal and figurative language, while other expressions test the boundary of compositionality. Multiword expressions of various types also create serious problems for computational linguistics, both in interpretation and generation of language.

Of course, there are existing theoretical (and in some cases computational) accounts of various types of multiword expression. We will review some of the most salient work briefly below. Plausible formal accounts can be found for the syntax and semantics of a considerable proportion of individual English constructions considered in isolation, although many taxing problems remain. Where current accounts are most lacking is in expressing connections and generalizations between multiword expressions and also their relationship to individual lexical items. Furthermore, insufficient attention has generally been paid to accounting for phrases that do not occur, and in elucidating the role of discourse information. Data resources are also lacking: the availability of corpora now enables us to look at natural occurrences of phrases, but we still lack detailed searchable lexicons. For instance, there is no way at the moment of finding all verbs that occur with the aspectual up particle (Levin (1993) does not cover verb particle combinations).

These issues concern linguists and computational linguists working within any framework who are interested in providing empirically adequate accounts or attempting to build broad-coverage lexicons. Accounts of multiword expressions also have a wider theoretical interest, in at least two ways. One is that properties of multiword expressions have often been used in arguments about more general syntactic properties. For instance, Nunberg et al (1994) review some of the ways in which various authors have used properties of idioms in support of theories. Similarly, examples involving compound nouns have been used to promote more general theories of semantic interpretation, especially with respect to the relative contribution of lexical semantics and discourse/pragmatics (e.g., Downing, 1977; Hobbs et al, 1993), but again more systematic exploration is called for. The second aspect of more general interest is the extension of formal devices apparently needed to handle multiword expressions, most controversially, the use of statistical information, which we will discuss below.

We propose to begin to address these issues systematically by developing a formal account of multiword expressions that has broad coverage over a range of naturally occurring data. We also intend to look at cross-linguistic aspects, as a way of validating and strengthening our account. Enough is already known about the formal representation issues that we feel confident that we can combine the theoretical work with a computational implementation. By implementing the account within the LinGO ERG and LKB system, we expect to be able to look at a large range of expressions with precision, to test generalizations and to provide a basis for a generally usable multiword lexical database. The implementation will also allow us to look at subtle frequency effects.

2 An overview of multiword expressions

In this section, we provide an overview of the main issues and of some previous work. We cannot give a full review here, but will concentrate on highlighting aspects that are most relevant to our proposal.

We distinguish between two broad categories of multiword expression, which we will describe using terminology adapted from Bauer (1983):

lexicalized phrases Phrases that have at least partially idiosyncratic syntax or semantics, or that contain 'words' that do not occur in isolation. The simplest cases might be described as words with spaces (e.g., ad hoc, of course) but we also take this category to include semantically idiosyncratic compound noun and adjective-noun combinations, idioms, verb-particle combinations, and various other constructions.[1]

institutionalized phrases Phrases that are syntactically and semantically predictable but that are used unexpectedly frequently (in a particular context). Examples in English include heavy toll, torrid affair, fine weather, enthusiasm evaporates/d, telephone booth (also telephone box in British English), kindle excitement, salt and pepper as well as standard similes (e.g., as dry as dust). The term collocation is sometimes used, but we reserve that to mean any statistically significant cooccurrence, including both lexicalized phrases as described above and compositional phrases that are frequent because of real world events. For instance, sell and house cooccur in sentences more often than would be predicted on the basis of the frequency of the individual words, but there is no reason to think that this is due to anything other than real world facts. Using our terminology, sell ... house is a collocation but not an institutionalized phrase.[2] Institutionalized phrases are not often discussed in formal linguistics, although there is some overlap with the concept of lexical function in Meaning Text Theory (e.g., Mel'čuk and Polguère, 1987). Lapata et al (1999) present evidence that suggests that the plausibility of adjective-noun combinations is strongly lexicalist and collocational in nature.

Lexicalized phrases require entries in the symbolic lexicon that capture their idiosyncratic properties. Institutionalized phrases, on the other hand, can be treated as only being statistically anomalous, and whether the statistical information is regarded as lexical is a matter of terminology (lexicalized phrases require both symbolic and statistical information). There is thus a clear implementational line between the two types, even though it is non-trivial to decide which category to assign to a phrase (or even a class of phrases). For instance, if it turns out that fine weather has a meaning more specific than would be expected on the basis of a compositional interpretation, then it would have to be treated as a lexicalized phrase. But 'extra' meaning may be accounted for by compositional processes: for instance, it may be reasonable to assume that fine weather implies that it is not raining simply because fine means something like good combined with cultural (rather than linguistic) assumptions about weather. The position that institutionalized phrases are linguistically relevant is perhaps somewhat controversial, but as we will see below, there are cases where this provides an interesting alternative to a purely symbolic account.

Multiword expressions can also be thought of as phrases, or phrasal patterns, that a native speaker will know and that should be accessible in an ideal dictionary, although no printed dictionary is likely to contain more than a fraction of the multiword expressions of the language. Jackendoff (1997:156) estimates that the number of multiword expressions in a speaker's lexicon is of the same order of magnitude as the number of simplex words, but it seems to us likely that this is an underestimate, even if we only include lexicalized phrases.

Syntactic variability Some fixed phrases, like ad hoc or of course, can be simply dealt with as words with spaces in. But generally, simple listing of multiword expressions as strings is inadequate, because many of them involve some degree of variability. There are idioms, such as kick the bucket, that are fixed apart from inflection[3] but these are probably the exceptions.

Verb-particle constructions, conjunctions like either ... or and so on, where material intervenes between the elements of the phrase, can be accounted for by means of a lexical selection mechanism where a sign associated with one word of the phrase selects for the other word(s). For instance, in the existing ERG, there is an entry for hand that subcategorizes for out in order to allow for word order variation (see also Sag, 1987).

[1] As used here, the term lexicalized phrase includes constructions that contain specific words such as 'X's way', and 'the Xer the Yer'. However, while we don't want to draw a hard line, our main interest here is in more lexically specific phrases.

[2] Examples like salt and pepper are cases where the coordination is presumably due to real world facts, but the ordering has to be regarded as conventionalized.

[3] Modulo examples like kick the proverbial bucket. However, this is a case where the claim that a construction is metalinguistic really appears valid, since the use of proverbial is presumably a comment on the language being used.

(1) Kim handed out the sweets to the kids

A lexical rule permutes the subcategorization list to allow:

(2) Kim handed the sweets out to the kids

This works if the syntactic relationship of the various parts of the phrase is fixed, as it indeed seems to be for verb particle constructions, but the mechanism runs into problems with some idioms, where the relationship between the words may be very flexible (see especially Nunberg, Sag and Wasow, 1994). Examples include let the cat out of the bag. This occurs in forms such as the (attested) example (British English speaker):

(3) That is a cat which has been a very long time coming out of its bag.

Examples such as this occur with considerable frequency in corpora (Riehemann, in preparation) and the argument that they are always metalinguistic is difficult to sustain (the utterer of the example above, for instance, denied conscious wordplay). Such examples are also problematic for an approach such as that of Erbach and Krenn (1994) that relies on the idiom having a head that can select for the other elements. Similarly, the TAG approach described by Abeillé (1990) faces difficulty here, although this is in most other respects probably the most satisfactory current computational account of idioms. Pulman (1993), Copestake (1994) and Riehemann (in preparation) all propose more semantically based accounts, but one question for such approaches is whether they can adequately account for the restrictions on variability of idioms.

The phenomenon of light verb noun combinations such as make a mistake, give a demo (but not ?do a mistake, ?make a demo) is similar to idioms in that, although an account where the verb selects its object is feasible, it seems at least as natural to treat the noun as determining the verb. Although such phrases are sometimes claimed to be idioms, this seems to be stretching the term too far: the noun is used in a normal sense, and the verb meaning appears to be bleached, rather than idiomatic.[4] Light verbs such as make in make a mistake can be used even when the object is a pronoun referring back to a previous utterance, as in 4a. It is not the case, however, that anything that could be regarded as an error can be made (see 4b).

(4) a. I've found another rounding error. He's made several of them in every calculation.

b. I've found another misspelled word. * He's made several of them in every paragraph.

Light verbs often do not translate consistently. For instance, the verb toru in Japanese is usually translated as take but sometimes corresponds to other verbs:

(5) kare-wa kyoka-o totta
he-top permission-acc took

He got permission

(6) kare-wa yasumi-o totta
he-top holiday-acc took

He took a holiday

There has been surprisingly little work on the formal syntax and semantics of these constructions in English, though has been some work that attempts to predict verb use (e.g., Wierzbicka, 1988:pp. 293–358). The comparable phenomenon in other languages has been better studied (e.g., Abeillé, 1988; Butt, 1995).

[4]The distinction is sometimes a little subtle. For instance, according to our assumptions, the phrase take care in good care was taken of the infants is not an idiom, because care is used in a standard sense, but Kim will take care of your problem involves an idiom, because care is idiosyncratic here. Note

(1) * Care was taken of your problem

.

Compositionality As discussed in detail by Nunberg, Sag and Wasow (1994), many idioms can be regarded as compositional, once the assumption is made that the words in the idioms have non-standard meanings. We will use the notation $\approx$ to relate an idiomatic word to a non-idiomatic near-equivalent (cf., translation equivalence). So, for instance, if we assume that in the idiom spill the beans, spill $\approx$ reveal and beans $\approx$ secret(s), the meaning of Kim spilled the beans, the Washington beans have been spilled can be derived by normal compositional processes. Of course, this still leaves the problem of ensuring that these meanings of spill and beans are only found in the context of the idiom. For other idioms, such as kick the bucket, and some other multiword expressions, it does not seem that the meaning of the phrase can usefully be divided between its components.

A somewhat different type of compositionality issue arises for compound nouns, such as cookie jar. These can be regarded as compositional, but only if some sort of relation is introduced as part of the meaning of the construction. One approach that has long been advocated in computational linguistics (e.g., Woods, 1972) and also in linguistics (e.g., Bauer, 1983) is that the relational meaning is completely underspecified by the construction (e.g., $[x][\text{cookie}[y] \wedge \text{jar}(x) \wedge \text{DUMMY}(y, x)]$, ignoring quantification) and then instantiated by constraints on the domain (or other contextual information). But this does not allow for systematic properties of classes of compound, such as stress patterns (Liberman and Sproat, 1992) and it results in overgeneration. There are classes of noun-noun compounds that are impossible in English, including, for instance, some types of compound involving relational nouns. An example is *spring beginning meaning beginning of spring, although the German Frühlingsanfang is normal. Systematic properties with respect to compound translation (e.g., Johnston and Busa, 1996) are also not fully explained by this account.

Semi-productivity Semi-productivity is apparent with some classes of multiword expressions in a similar way to derivational morphology. Besides the impossible compound patterns, mentioned above, compound schemata range from the non-productive (e.g., the verb-noun pattern exemplified by pickpocket), to the almost fully productive (e.g., made-of), with many schemata being intermediate (e.g., has-part: 4-door car is acceptable but the apparently similar *sunroof car is not). Similarly, there are some cases where combination with a particle apparently should be regarded as productive within a class of verbs (e.g., sweep up, mop up, vacuum up, hoover up), but there are also many cases where there is far less regularity.

Generalizations We can distinguish at least two dimensions on which we want to express generalizations. The first is the commonalities in behaviour among members of classes, which can be represented using devices such as the HPSG type hierarchy, as we mentioned in our description of our current project. The second type of generalization involves common elements in multiword expressions: e.g., the aspectual up in tear up, stir up etc; hit meaning (something like) reach in hit the headlines, hit the hay. Another example is the use of light in various idioms that reflect the 'seeing as understanding' metaphor (Sweetser 1990). In general, many idiom patterns may involve underlying metaphors (e.g., Lakoff and Johnson, 1980).

There are also cases that might be better regarded as multiword expressions with variable components: e.g. the idiom throw someone to the wolves/lions/dogs. Similarly, many lexicalized phrases can be regarded as being connected to lexical entries for simplex words. For instance, tea towel is lexicalized (at least in British English), since its meaning has nothing much to do with tea, but regarding it simply as a word with a space seems inadequate, since there is a connection with towel.

Computational aspects Conventional approaches to parsing have problems with multiword expressions analogous to those just discussed in a theoretical context: that is, it is necessary to allow variability and productivity without overgenerating. Generalization is important in building a maintainable grammar and lexicon. Unexpected multiword expressions are a major cause of failure in broad coverage grammars: many phrases cannot be parsed unless they have lexical entries as multiwords. However, in many ways, cases where a phrase can also be analysed compositionally are worse, because the wrong meaning will be produced but the failure is difficult to detect automatically. In a translation system, for instance, this will result in an inappropriate literal translation.

Institutionalized phrases cause particular problems for generation, because the use of a non-standard phrase, or of an institutionalized phrase in an unusual context, while grammatical, may be jarring. For instance, conjunction order is semi-fixed in examples like knife and fork and hot and cold and to invert the order has a marked effect. Similarly, while beautiful weather and foul weather are usual, pretty weather and putrid weather are relatively strange expressions. Their generation should not be completely blocked, but should be avoided except in a context where a marked phrase is appropriate.[5]

A general issue in interpretation of multiword expressions is how to ensure that the idiosyncratic interpretation is only chosen when appropriate. For instance, as well as has the idiosyncratic interpretation in addition to but also the productive interpretation in an equally good manner. So the following example is ambiguous:

(7) I could do it as well as you.

This case is actually somewhat unusual in that the two interpretations are both common. More often, the idiosyncratic interpretation is much more frequent. However, generally the productive interpretation is still available, so the heuristic of always choosing the multiword expression will not work in general.

Statistical approaches to processing, especially those which adopt lexicalist models, have at least an implicit notion of lexicalized and institutionalized phrases. The language models conventionally used in speech recognition are perhaps the best examples of this, since statistically significant collocations are often reflected in bigram or trigram models. Lexicalized models have also been used in learning more syntactically interesting grammars (e.g., Collins, 1997). Bod (1998) uses the idiosyncracy of multiword expressions to argue that the Data-Oriented Parsing (DOP) model is in principle preferable to an approach that adds stochastic information to a conventional grammar. The practical problem that arises, however, is that enormous amounts of data are needed to build good models for anything other than the most frequent words. Institutionalized phrases complicate this, because a word that is part of such a phrase may well have very different properties in that context. Hence finding ways to generalize over multiword expressions is potentially as relevant as it is to traditional approaches to parsing, even if the methodology is very different.

REFERENCES CITED

Anne Abeillé (1988) 'Light verb constructions and extraction out of NP in a tree adjoining grammar', Papers of the 24th regional meeting of the Chicago Linguistic Society, Chicago.

Anne Abeillé (1990) 'Lexical and Syntactic rules in a Tree Adjoining Grammar', Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL-90), Pittsburgh, pp. 292–298.

Nicholas Asher (1993) Reference to Abstract Objects in Discourse, Kluwer Academic Publishers.

Ken Bame (1999) 'Aspectual and resultative verb-particle constructions with up', Handout of talk given at Ohio State University Graduate Linguistics Student Colloqium.

Laurie Bauer (1983) English word-formation, Cambridge University Press, Cambridge, England.

Emily Bender and Daniel Flickinger (1999a) 'Peripheral constructions and core phenomena' in Andreas Kathol, Jean-Pierre Koenig, Gert Webelhuth (ed.), Lexical and Constructional Aspects of Linguistic Explanation, CSLI Publications, pp. 199–214.

Emily Bender and Daniel Flickinger (1999b) 'Diachronic evidence for extended Argument Structure' in Gosse Bouma, Erhard Hinrichs, Geert-Jan Kruijff and Richard Oehrle (ed.), Constraints and Resources in Natural Language Syntax and Semantics, CSLI Publications.

Emily Bender and Ivan A. Sag (1999) 'Incorporating Contracted Auxiliaries in English', Paper presented at the Sixth Annual Conference on Head-Driven Phrase Structure Grammar (HPSG-99), University of Edinburgh.

Rens Bod (1998) Beyond grammar: an experiencer-based theory of language, CSLI Publications.

Gosse Bouma, Daniel Flickinger, and Frank van Eynde (2000) 'Constraint-based lexica' in Frank van Eynde and Daffyd Gibbon (ed.), Lexicon Development for Speech and Language Processing, Kluwer: Dordrecht.

[5]Statistical approaches to generation that make use of bigrams (e.g., Langkilde and Knight, 1998) tend to get these expressions right.

Gosse Bouma, Robert Malouf, and Ivan A. Sag (in press) 'Satisfying Constraints on Extraction and Adjunction', Natural Language and Linguistic Theory.

Ted Briscoe and Ann Copestake (1999) 'Lexical rules in constraint-based grammars', Computational Linguistics, 25:4, 487–526.

Miriam Butt (1995) The structure of complex predicates in Urdu, CSLI Publications.

Bob Carpenter (1992) The Logic of Typed Feature Structures, Cambridge University Press, Cambridge, England.

John Carroll, Ann Copestake, Daniel Flickinger and Victor Poznanski (1999) 'An efficient chart generator for (Semi-)lexicalist grammars', Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99), Toulouse, pp. 86–95.

Michael Collins (1997) 'Three generative, lexicalised models for statistical parsing', Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL 97), Madrid, pp. 16–23.

Ann Copestake (1992) 'The Representation of Lexical Semantic Information', D.Phil. dissertation, University of Sussex, Brighton, England. Cognitive Science Research Paper CSRP 280.

Ann Copestake (1994) 'Representing idioms', Paper presented at the HPSG Conference, Copenhagen.

Ann Copestake (1997) 'Augmented and alternative NLP techniques for augmentative and alternative communication', Proceedings of the ACL workshop on Natural Language Processing for Communication Aids, Madrid, pp. 37–42.

Ann Copestake (1999) 'The (new) LKB system', http://www-csli.stanford.edu/∼aac/newdoc.pdf.

Ann Copestake (in press) 'Definitions of Typed Feature Structures', Natural Language Engineering (appendix to special issue on efficient processing with HPSG), 6(1).

Ann Copestake (in preparation) Implementing Typed Feature Structure Grammars, CSLI Publications.

Ann Copestake and Daniel Flickinger (1998) 'Enriched Language Models for Flexible Generation in AAC systems', Proceedings of the 13th Annual Conference, Technology and Persons with Disabilities (CSUN), Los Angeles, CA.

Ann Copestake and Daniel Flickinger (1999) 'Evaluation of NLP technology for AAC using logged data' in Filip Loncke, John Clibbens, Helen Arvidson and Lyle Lloyd (ed.), Augmentative and Alternative Communication: new directions in research and practice, Whurr Publishers, London, pp. 123–132.

Ann Copestake and Daniel Flickinger (2000) 'An open-source grammar development environment and broad-coverage English grammar using HPSG', Proceedings of the Second conference on Language Resources and Evaluation (LREC-2000), Athens, Greece.

Ann Copestake, Daniel Flickinger, Ivan Sag and Carl Pollard (1999) 'Minimal Recursion Semantics: An introduction', `http://www-csli.stanford.edu/∼aac/papers/newmrs.ps`.

Ann Copestake and Alex Lascarides (1997) 'Integrating symbolic and statistical representations: the lexicon-pragmatics interface', Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL 97), Madrid, pp. 136–143.

Ann Copestake and Alex Lascarides (1998) 'Resolving Underspecified Values with Discourse Information', Paper presented at the workshop on models of underspecification and the representation of meaning, Bad Teinach, Germany.

Pamela Downing (1977) 'On the Creation and Use of English Compound Nouns', Language, 53(4), 810–842.

Gregor Erbach and Brigitte Krenn (1994) 'Idioms and Support-Verb Constructions in HPSG' in John Nerbonne, Klaus Netter and Carl Pollard (ed.), German in Head-driven Phrase Structure Grammar, CSLI Lecture Notes.

Christiane Fellbaum (1993) 'The determiner in English idioms' in C. Cacciari and P. Tabossi (ed.), Idioms: Processing, Structure, and Interpretation, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 271–296.

Charles Fillmore, Paul Kay and Mary C. O'Connor (1988) 'Regularity and idiomaticity in grammatical constructions', Language, 64, 501–538.

Charles Fillmore and Paul Kay (to appear) Construction Grammar, CSLI Publications.

Daniel Flickinger (1987) 'Lexical Rules in the Hierarchical Lexicon', PhD dissertation, Stanford University, Stanford, CA.

Daniel Flickinger (in press) 'On building a more efficient grammar by exploiting types', Journal of Natural Language Engineering, (Special Issue on Efficient Processing with HPSG), 6(1).

Daniel Flickinger, Stephan Oepen, Hans Uszkoreit, and Jun-ichi Tsujii (editors) (in press) 'Journal of Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG)', Cambridge University Press.

Jonathan Ginzburg and Ivan A. Sag (1999) 'Constructional Ambiguity in Conversation', Proceedings of the Proceedings of the 12th Amsterdam Colloquium, University of Amsterdam.

Jonathan Ginzburg and Ivan A. Sag (2000) 'In Situ Wh-Interrogatives', Paper presented at the 36th Regional Meeting of the Chicago Linguistic Society, Chicago, Illinois.

Jonathan Ginzburg and Ivan A. Sag (in preparation) 'English Interrogative Constructions', CSLI Publications.

Eirik Hektoen (1997) 'Probabilistic parse selection based on semantic cooccurrences', Proceedings of the 5th International workshop on parsing technologies (IWPT-97), MIT, Cambridge, Mass., pp. 113–122.

Jerry Hobbs, Mark Stickel, Doug Appelt and Paul Martin (1993) 'Interpretation as Abduction', Artificial Intelligence, 63.1, 69–142.

Ray Jackendoff (1997) The Architecture of the Language Faculty, MIT Press.

Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi and Stefan Riezler (1999) 'Estimators for stochastic "Unification-based" grammars', Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), University of Maryland, USA, pp. 535–541.

Michael Johnston and Frederica Busa (1996) 'Qualia structure and the compositional interpretation of compounds', Proceedings of the ACL SIGLEX workshop on breadth and depth of semantic lexicons, Santa Cruz, CA.

Martin Kay (1996) 'Chart Generation', Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), Santa Cruz, CA, pp. 200–204.

Bernd Kiefer, Hans-Ulrich Krieger, John Carroll and Rob Malouf (1999) 'A Bag of Useful Techniques for Efficient and Robust Parsing', Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), University of Maryland, USA, pp. 473–480.

Jong-Bok Kim and Ivan A. Sag (in preparation) 'Negation without Head-Movement.', Manuscript under revision for Natural Language and Linguistic Theory.

George Lakoff and Mark Johnson (1980) Metaphors we live by, University of Chicago Press, Chicago.

Irene Langkilde and Kevin Knight (1998) 'The practical value of n-grams in generation', Proceedings of the Ninth international workshop on Natural Language Generation, Niagara-on-the-Lake, Ontaio, Canada, pp. 248–255.

Maria Lapata, Scott McDonald and Frank Keller (1999) 'Determinants of adjective-noun plausibility', Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 99), Bergen, pp. 30–36.

Alex Lascarides and Nicholas Asher (1993) 'Temporal Interpretation, Discourse Relations and Common Sense Entailment', Linguistics and Philosophy, 16, 437–493.

Alex Lascarides and Ann Copestake (1999) 'Default representation in constraint-based frameworks', Computational Linguistics, 25:1, 55–106.

Beth Levin (1992) Towards a Lexical Organization of English Verbs, University of Chicago Press, Chicago, IL.

Mark Liberman and Richard Sproat (1992) 'The stress and structure of modified noun phrases in English' in Ivan A. Sag and Anna Szabolsci (ed.), Lexical matters, CSLI Publications, pp. 131–182.

Dekang Lin (1999) 'Automatic identification of non-compositional phrases', Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), University of Maryland, USA, pp. 317–324.

Rob Malouf (in press) 'The order of prenominal adjectives in natural language generation', Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong.

Rob Malouf, John Carroll and Ann Copestake (in press) 'Efficient feature structure operations without compilation', Journal of Natural Language Engineering, (Special Issue on Efficient Processing with HPSG), 6(1).

Igor Mel'čuk and Alain Polguère (1987) 'A formal lexicon in Meaning-Text Theory (or how to do lexica with words)', Computational Linguistics, 13:3–4, 261–275.

Guido Minnen, Francis Bond and Ann Copestake (in press) 'Memory-based learning for article generation', Proceedings of the Fourth Computational Natural Language Learning Workshop (CoNLL-2000), Lisbon.

Günter Neumann (1997) 'Applying explanation-based learning to control and speeding-up natural language generation', Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL 97), Madrid, pp. 214–221.

Geoffrey Nunberg, Ivan A. Sag and Thomas Wasow (1994) 'Idioms', Language, 70, 491–538.

Stephan Oepen and John Carroll (2000) 'Ambiguity Packing in Constraint-based Parsing. Practical Results', Proceedings of the First Conference of the North American Chapter of the ACL (NACL 2000), Seattle, WA.

Stephan Oepen and John Carroll (in press) 'Performance Profiling for Parser Engineering', Journal of Natural Language Engineering (Special Issue on Efficient Processing with HPSG), 6(1).

Stephan Oepen and Daniel Flickinger (1998) 'Towards Systematic Grammar Profiling. Test Suite Technology Ten Years After', Journal of Computer Speech and Language: Special Issue on Evaluation, 12 (4), 411-437.

Stephan Oepen, Dan Flickinger, Hans Uszkoreit, and Jun-ichi Tsujii (editors) (in preparation) Efficiency in Unification-Based Processing, CSLI Publications.

Carl Pollard and Ivan A. Sag (1987) An information-based approach to syntax and semantics: Volume 1 fundamentals, CSLI Lecture Notes 13, CSLI Publications, Stanford CA.

Carl Pollard and Ivan A. Sag (1994) Head-driven Phrase Structure Grammar, The University of Chicago Press, Chicago and CSLI, Stanford.

Steven G. Pulman (1993) 'The Recognition and Interpretation of Idioms' in C. Cacciari and P. Tabossi (ed.), Idioms: Processing, Structure, and Interpretation, Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 249–270.

James Pustejovsky (1995) The Generative Lexicon, MIT Press.

Susanne Z. Riehemann (in preparation) 'A constructional approach to idioms and word formation', PhD Thesis, Stanford University.

Susanne Z. Riehemann and Emily Bender (1999) 'Absolute Constructions: On the Distribution of Predicative Idioms', Proceedings of the 18th West Coast Conference on Formal Linguistics (WCCFL-18), Tuscon, AZ, Cascadilla Press.

Ivan A. Sag (1987) 'Subcategorization, Grammatical Hierarchy and Linear Precedence.' in Geoffrey Huck and Almerindo Ojeda, eds. (ed.), Discontinuous Constituency; Syntax and Semantics Volume 20., Academic Press, pp. 303–340.

Ivan A. Sag (1997) 'English Relative Clause Constructions', Journal of Linguistics, 33(2), 431–484.

Ivan A. Sag (2000) 'Rules and Exceptions in the English Auxiliary System', Proceedings of the HPSG99 – Seventh Annual Conference on Head-Driven Phrase Structure Grammar, University of California, Berkeley.

Ivan A. Sag and Thomas Wasow (1999) Syntactic Theory — a formal introduction, CSLI Publications, Stanford CA (also distributed by Cambridge University Press).

Sayori Shimohata, Toshiyuki Sugio and Junji Nagata (1997) 'Retrieving collocations by co-occurrences and word order constraints', Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL 97), Madrid, pp. 476–481.

Jeffrey D. Smith (1999) 'English Number Names in HPSG' in Andreas Kathol, Jean-Pierre Koenig, Gert Webelhuth (ed.), Lexical and Constructional Aspects of Linguistic Explanation, CSLI Publications, pp. 145–160.

Eve Sweetser (1990) From etymology to pragmatics, Cambridge University Press, Cambridge, England.

Anna Wierzbicka (1988) The semantics of grammar, John Benjamins, Amsterdam.
William Woods (1972) The Lunar sciences natural language information system, Final Report, Bolt, Beranek and Newman, Cambridge, MA.