BAAL Corpus Linguistics SIG / OTA Workshop: *Identifying and Researching Multi-Word Units*

Oxford University Computing Services

21st April 2005

Time	Speaker	Title
9:30	Coffee and registration	
10-11am	Pernilla Danielsson	**More than words: a study of what constitutes a multi-word unit**
11-12	William Fletcher	**Real-Time Identification of MWE Candidates in Data from the BNC and the Web**
12-12:30	Pete Whitelock	**Dependency parsing and comprehensive collocation extraction**
12:30-1:30	Lunch	
1:30-2:30	Paul Rayson	**Right from the word go: identifying multi-word-expressions for semantic tagging**
2:30-3	Software Demonstrations	
3-3:30	Wenzhong Li	**Extracting Multi-word Units — An analysis of the word clusters in China's English news articles**
3:30-4:00	Tea	
4-4:30	Nicholas Groom	**Using salient grammatical words to identify and analyse multi-word units**
4:30-5	Violeta Seretan	**Beyond Word Pairs: Extracting Long Multi-Word Units from Parsed Text Using a Method of Pair Composition**

Organised by:
Susan Hunston s.e.hunston@bham.ac.uk
Paul Thompson p.a.thompson@reading.ac.uk
Martin Wynne martin.wynne@oucs.ox.ac.uk

http://corpus-sig-baal.org.uk/
http://ota.ox.ac.uk/
http://www.ahds.ac.uk/litlangling/

ahds literature, languages and linguistics

Abstracts

More than words: a study of what constitutes a multi-word unit

Pernilla Danielsson, University of Birmingham a.p.danielsson@bham.ac.uk

This talk will focus on what we perceive to be multi-word units in texts. As the attention has grown on identifying larger units in language, it has become apparent that we have yet to reach a definition for what can be seen as a valid multi-word unit. What makes a multi-word unit interesting? Where does the unit begin and end? Can we still talk about one and the same unit (the prototypical unit) even if there is slight modification within it? The talk will include findings from several studies, discussing both automatic identification, human identification of multi-word units and how looking at larger units in language will affect regularities of the distribution of words.

Real-Time Identification of MWE Candidates in Data from the BNC and the Web

William Fletcher, US Naval Academy, on sabbatical at Radboud University Nijmegen fletcher@kwicfinder.com

This paper compares techniques for rapid automatic identification of MWE candidates -- here a cover term for multi-word units and salient collocations -- for application to dynamically-generated datasets where speed is crucial to user satisfaction. It reports on work-in-progress for several projects: the "Phrases in English" online n-gram database based on the BNC (http://pie.usna.edu/), to be augmented with data from MICASE and the ANC; development of the tools KWiCFinder web concordancer and kfNgram n-gram analysis tool (free downloads from http://kwicfinder.com/); and a nascent effort to establish a consortium for "Web as Corpus". Identification of MWEs will support various goals such as describing subcorpora in large linguistic databases, identifying content domain and text-type for Web documents, and compiling reference and instructional materials for language learners and professionals. In this paper two lexical association metrics will be compared in detail: Mutual Information and Mutual Rank Ratio, developed by Paul Deane of the Educational Testing Service. Practical aspects of the techniques will be stressed -- speed and scalability -- rather than their effectiveness in terms of precision and recall or their theoretical justification.

Dependency parsing and comprehensive collocation extraction

Pete Whitelock, Sharp Laboratories pete@sharp.co.uk

This paper describes the design of a flexible dependency parser and its use to bootstrap a comprehensive description of English word combinations and their collocational strengths. A wide-coverage account of this sort has many applications: enhancing the accuracy of parsing and thus in turn the collocational data itself; improving the choice of words in (human and machine) translation and/or generation; detecting and correcting errors in written English.

In a similar way to researchers such as Lin and Kilgarriff, and unlike earlier attempts to extract collocations based on continuous n-grams or text windows, we assign collocational strength to connected regions of dependency structure. Our novel parsing algorithm first computes a set of potential dependency links based on fine-grained pos-tagging. We then extract a consistent maximal subset using link strength scores based on tag frequencies, word separations and, as the bootstrapping progresses, collocation strengths.

We dependency-parsed 80m. words of the BNC, computed the strength of the 12m. pairs and triples of words that occurred more than once, repaired the parse using these values and recomputed the collocation strengths. The result is comprehensive in that we detect not only true

(strong/idiomatic/non-compositional) collocations, but can assign a strength to every syntactic combination, offering the possibility of high coverage error detection.
The resulting data can be found at http://www.sle.sharp.co.uk/JustTheWord. Syntactic structure and synonym sets are used to organise the presentation of the strong collocates of a given word. The system can also offer suggestions for the improvement of improbable word combinations.

Right from the word go: identifying multi-word-expressions for semantic tagging

Paul Rayson, Lancaster University **paul@comp.lancs.ac.uk**

In this talk I will describe our approach to the computational identification of multi-word-expressions (MWE) in corpora. The work stems from a larger effort to construct a semantic field tagger for English. I will describe what types of MWEs we target, how they are identified and what potential for variation is allowed. There are two distinct tasks in our approach (i) identification of candidate MWEs (ii) classification of the MWEs according to our semantic field taxonomy. I will focus on the identification task and describe the hydrid of rule-based and statistical techniques used to find low frequency MWEs and previously unseen MWEs. We are currently extending these techniques to Finnish and Russian.

Extracting the Multi-word Units – an analysis of the word clusters in China’s English news articles

Wenzhong Li, University of Central England **li.wzhong@gmail.com**

The research is designed as a pilot study to investigate the typical patterns of word clusters in China’s English news articles and to see how words are clustered to express things uniquely Chinese. It has been discovered that an average of 14.53% of all the n-word clusters under investigation are found unusually frequent of their occurrence in China’s news articles; the 4, 5, 6 word clusters are the most frequent ones used. The statistics shows that the longer the clusters, the greater specificity of meaning, the more fixed of the cluster patterns. When the size of the word clusters increases, the chance of clusters as pre-fabricated chunks standing alone also increases steadily. It is also discovered that the longer the clusters, the more oriented to politics they are. It has been concluded that word clusters of varied length are prevalent in China English news articles. And such clusters contribute a lot to the formation of fixed expressions, phrases, and collocations that are unique in China context of English use. Longer clusters form mini texts that convey messages representing to a great extent the social conventions, political ideology, and economic inventories uniquely Chinese characterized. The invented words together with the ready-made ones are used with marked meanings and specific intentions. The most explicit features of China English are found in word combinations and texts, as in which the complete message is conveyed.

Using salient grammatical words to identify and analyse multi-word units

Nicholas Groom, University of Birmingham **Nick@nicholasgroom.fsnet.co.uk**

This talk focuses on methodological and theoretical issues arising from an AHRB-funded doctoral research project on phraseological variation in academic writing in the Humanities. This study follows the groundbreaking work of Gledhill (e.g. 2000a, 2000b) in that it uses statistically salient grammatical words as probes for phraseological sequences (i.e. multi-word units) in natural language corpus data.
In this talk, I will describe this procedure in detail, and present some examples of the kinds of data that it yields. My broad aim will be to highlight the particular strengths and limitations of this 'bottom-up' or lexically-driven approach to the identification and analysis of multi-word units, and to assess the extent to which it complements, or even constitutes a viable alternative to, the increasingly widespread use of n-grams in corpus-based phraseological research.

References:

Gledhill, C. (2000a). *Collocations in science writing.* Tübingen: Gunter Narr.

Gledhill, C. (2000b). The discourse function of collocation in research article introductions. *English for Specific Purposes* 19: 115-135.

Beyond Word Pairs: Extracting Long Multi-Word Units from Parsed Text Using a Method of Pair Composition

Violeta Seretan, University of Geneva Violeta.Seretan@lettres.unige.ch

We report on an implemented method of phraseological unit extraction specifically oriented towards the detection of flexible, well-formed expressions made up of more than two words - e.g., collocations or phrasal templates, such as: "fierce battle rage", "play a leading role in", "turn a blind eye to", "weapon of mass destruction by force", "country take the helm for _ months" (actual extraction results, showing lexemes rather than inflected forms).

While there exist a wide range of methods and tools dedicated to the automatic identification of word pairs (see for instance (Manning and Schütze, 1999) for a review of statistical methods employed for collocation extraction), the methodologies for extracting longer units are considerably less developed. The systems developed generally limit to rigid sequences of adjacent words of length 6 or 7 (Choueka, 1983; Dias, 2003); as for flexible combinations, usually no more than 3 words are considered (Zinsmeister and Heid, 2003).

We implemented an extraction method suitable for long and flexible units, which is based on syntactic parsing and is not affected by combinatorial explosion. We identify this kind of units by processing the results of Fips (a syntactic parser based on GB theory). In the first step, all the word pairs in given syntactic configurations are extracted. Secondly, a procedure of lexical chain building is applied, that joins the pairs sharing common items. We follow a general algorithm of "bigram composition" (Seretan et al., 2003) in order to build up word chains of unrestricted length, which constitute unit candidates.

A preliminary experiment led to the identification of flexible unit candidates of up to 9 words, and proved the potential of this method to deal with phrases showing a high degree of syntactic variability. A dedicated concordance tool allows for the ease visualization of the extracted candidates (all the various instances of a unit are grouped under the same base form) and for their possible validation by a lexicographer.

References

Choueka, Yaacov, S.T. Klein, and E. Neuwitz, 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4(1):34-38.

Dias, Gaël, 2003. Multiword unit hybrid extraction. In *Proceedings of the Workshop on Multiword Expressions at the 41th Annual Meeting of the Association for Computational Linguistics* (ACL'03) pages 41-48, Sapporo, Japan.

Manning, Christopher and Heinrich Schütze, 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Seretan, Violeta, Luka Nerima, and Eric Wehrli, 2003. Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the Fourth International Conference on Recent Advances in NLP* (RANLP-2003) pages 424-431, Borovets, Bulgaria.

Zinsmeister, Heike and Ulrich Heid, 2003. Significant triples: Adjective+ noun+verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research* (Complex 2003), Budapest.

Participants

Lina Adinolfi	The Open University
Wendy Anderson	University of Glasgow
Kevin Armstrong	University of Leicester
Fiona Barker	Cambridge ESOL
Ylva Berglund	Oxford University
Jill Bowie	Reading University
Lou Burnard	Oxford University
Maggie Charles	Oxford University
Janet Cotterill	Cardiff University
David Cram	Oxford University
James Cummings	Oxford University
Graham Cunningham	Oxford University
Pernilla Danielsson	University of Birmingham
Jarle Ebeling	Oxford University
Philip Edmonds	Sharp Labs. of Europe Ltd.
Bill Fletcher	US Naval Academy
Richard Forsyth	University of Warwick
Nicholas Groom	University of Birmingham
Susan Hunston	University of Birmingham
Frank Keenan	Oxford University Press
Chris Kennedy	University of Birmingham
Judith Kennedy	University of Warwick
Deborah Keogh	Trinity College Dublin
Hui-Ling Lang	Ming Chuan University
Maria Leedham	Oxford Brookes University
Oliver Mason	University of Birmingham
John McKenny	Northumbria University
Peet Morris	Oxford University
Hilary Nesi	University of Warwick
Richard Poole	Oxford University Press
Victor Poznanski	Sharp Labs. of Europe Ltd.
Paul Rayson	Lancaster University
Alison Sealey	Aston University
Violeta Seretan	University of Geneva
Tania Shepherd	Universidade do Estado do Rio de Janeiro
Greg Simpson	Oxford University
Dominic Smith	Birmingham University
Paul Thompson	Reading University
Samuel Tomblin	Cardiff University
Robert Vanderplank	Oxford University
Pete Whitelock	Sharp Labs. of Europe Ltd.
Rowan Wilson	Oxford University
Martin Wynne	Oxford University
Wenzhong Li Henan	Normal University, Xinxiang, P.R. China