Extraction Methods - description[1]

Irina Dahlmann
(November 2007)

Extraction Methods - description

‘Clusters’ (WordSmith Tools 4.0)

‘Multi-word expressions’ (MWEs - WMatrix2)

‘Clusters’ (WordSmith Tools 4.0)

Background

WordSmith Tools is a lexical analysis software developed by Mike Scott in Liverpool.[2] The function which is interesting from a multi-word unit perspective is the extraction of ‘clusters’, which - within WordSmith Tools 4.0 - is defined as ‘a group of words which follow each other in a text’[3].

[1] Published on https://www.nottingham.ac.uk/english/research/cral/doku.php?id=projects:slsf; Project website for 'Second language speech fluency (SLSF) and the role of pauses in automatically extracted multi-word units (MWUs)'

[2] http://www.lexically.net/wordsmith/index.html

[3] http://www.lexically.net/wordsmith/index.html, entry ‘cluster definition’

Clusters

> *‘Clusters are words which are found repeatedly together in each others' company, in sequence. They represent a tighter relationship than collocates, more like multi-word units or groups or phrases. (I call them clusters because groups and phrases already have uses in grammar and because simply being found together in software doesn't guarantee they are true multi-word units.) Biber calls them "lexical bundles".’*
>
> *(http://www.lexically.net/downloads/version4/html/index.html*
>
> *last accessed 10-06-2007)*

Another frequently used term is ‘n-gram’, which refers to word clusters of a certain lengths of words (n words, i.e. 2-grams consist of 2 subsequent words; 3-grams consist of 3 subsequent words etc.).

Search mechanism for clusters in WordSmith Tools 4.0

The search mechanism as described by Mike Scott (see below) allows for a search of recurrent and *continuous* sequences of words within a given data set:

Search mechanism

‘Suppose your text begins like this:

Once upon a time, there was a beautiful princess. She snored. But the prince didn't.

If you've chosen **2-word clusters** the text will be split up as follows:

Once upon

upon a

a time

(note ***not*** *"time there" because of the comma)*

there was (etc.)

With a **three-word cluster** setting it would be split up like this:

> *Once upon a*
>
> *upon a time*
>
> *there was a*
>
> *was a beautiful*
>
> *a beautiful princess*
>
> *But the prince*
>
> *the prince didn't*
>
> *(etc.)*
>
> *That is, each n-word cluster will be stored, if it reaches n words in length, up to a punctuation boundary, marked by* ***;,.!?*** *(It seems reasonable to suppose that a cluster does not cross clause boundaries and these punctuation symbols help mark clause boundaries.)*

(http://www.lexically.net/downloads/version4/html/index.html

last accessed 10-06-2007)

In other words, an n-word cluster, n-gram or lexical bundle will be recognised, counted and stored if it occurs repetitively in the actual corpus used. In our case, our own two corpora ENSIC and NICLE(s) thus are the sole basis for the search. Only if certain strings are used repeatedly in this particular data set, they will be counted as clusters. This, for instance, is one of the main differences to the extraction with WMatrix2: WMatrix2 is based on a data base of currently 18,922 '**m**ulti-**w**ord **e**xpression' (MWE) templates[4] which the corpus in question is tested against. This means for example that WMatrix2 can also find MWEs, which only occur once in the corpus, as it is not dependent on the corpus data alone but draws on a predefined set of MWEs. (see also section on WMatrix2)

[4] As to 10/06/2007 (New entries are still added to the database; however, there have only been 190 additions between the 18,732 MWE templates reported by Piao et al. in 2005 and the current 18,922 templates.)

Suitability of WordSmith Tools' 'cluster' search for NICLE(s) and ENSIC

WordSmith Tools uses a simple and mechanic technique of searching texts for recurrent clusters. It is a statistical approach based on frequency and co-occurrence (Scott et al. 2003)

One of the great advantages of generating clusters with WordSmith Tools is that it is not based on preconceived ideas of what a cluster is good. This is particularly important when working with learner data. Learners may use their own idiosyncratic sets of MWUs or MWUs which not necessarily confirm with native speaker usage (but may well be used very frequently within a group of learners) and they are therefore not reflected in data bases which are generally based on native English.

On the other hand, only because words in a cluster occur together these clusters are not automatically 'true multi-word units' as the co-occurrence of words does not necessarily imply a grammatical relation between the words. 'Though' - as Scott remarks - 'clusters often do match phrases or idioms'.[5]

However, even for native speech it might generate results of clusters which one normally would not regard as a MWU (by intuition, lack of grammatical relation etc, such as '*I I I*', '*but I*', '*er I think*', '*I think I*'); but if those clusters occur very frequently together it might be worth investigating them further (for instance with pause analysis) to find out whether the traditional view of MWUs might need to be broadened slightly. good

A further drawback from a multi-word unit perspective is that this approach does not allow for variations in sequences or discontinuous sequences and are unable to deal with low-frequency MWEs, which is particularly problematic considering that

> *'the majority of the words in most corpora have low frequencies, occurring only once or twice. This means that a major part of true multiword expressions are left out by statistical approaches'. (Scott et al. 2003)*

However, our approach seeks to find *patterns* of usage (pause phenomena in particular) of several individual MWEs, therefore frequent occurrence in the corpus is one of the criteria for the MWEs candidates which will be studied in more depths. Low frequent MWEs can be neglected for this project. They are nevertheless included in the result lists as they may give a fuller picture of the used corpora and may be interesting from other research perspectives.

‘Multi-word expressions’ (MWEs - WMatrix2)

The software tool for corpus analysis and corpus comparison WMatrix2 has been developed by Paul Rayson at Lancaster University[6]. Wmatrix2 is also the web interface to the USAS and CLAWS corpus annotation tools[7], which are both utilised in the extraction of MWEs. For more detailed information on USAS and CLAWS consult Piao et al. (2005a, 2005b; 2003) and the UCREL website[8].

The Lancaster semantic lexicon[9]

The extraction of MWEs by WMatrix is based on a large semantic lexicon, the *Lancaster semantic lexicon*. It contains two kinds of entries, single word entries and multi-word expression (MWE) templates, and maps lexemes and MWEs to their potential semantic categories. The taxonomy contains 21 major semantic fields, such as *Time, Education, General and abstract terms, Numbers and measurements etc.* which are further divided into 232 sub-categories. Entries can be assigned multiple senses and belong to different semantic fields. In the analysis procedure, these are ‘disambiguated according to their context in use’ and assigned the ‘correct’ sense in the tagging process via a semantic field analysis using the semantic tagger USAS (the **U**CREL **S**emantic **A**nalysis **S**ystem).

[5] http://www.lexically.net/wordsmith/index.html
[6] Rayson 2001-7, http://ucrel.lancs.ac.uk/wmatrix2.html
[7] http://ucrel.lancs.ac.uk/wmatrix2.html
[8] http://www.comp.lancs.ac.uk/ucrel/
[9] All information on the Lancaster semantic Lexicon is based on Piao et al. 2005a.

MWE entries

The following description is taken from Piao et al. 2005a.

Each entry together with its part of speech (POS) annotation is mapped to a semantic category. ‘For example as shown in Fig. 2, the word “iron” is mapped to the category of [*S1.2.5+: Toughness, Strong/Wea*k] when it is used as an adjective, to the category of [O1.1: Object/Substance], [*B4: Cleaning and Personal Care*] and [*O2:material*] when used as a noun, and to the category of [*B4: Cleaning and Personal Care*] when used as a verb.

Sample of single word entries:

iron	JJ	S1.2.5+
iron	NN1	O1.1 B4/O2 O2
iron	VV0	B4
ironic	JJ	X2.6-
ironical	JJ	X2.6-

The entries in the MWE sub-lexicon have similar structures as the single word counterpart but the key words are replaced by MWEs. Here, the combination of constituent words of each MEW depicts a single semantic entity, and thus are mapped to semantic category/ies together. For example, the MWE “life expectancy” is mapped to the categories of [*T3: Time/Age*] and [*X2.6: Expect*]. In addition, MWEs that share similar structures and belong to the same semantic space are transcribed as templates using a simplified form of a regular expression. For example, the template [*ing_NN1 machine*_NN*] represents a set of MWEs including “washing machine/s”, “vending machine/s” etc. [Also, asterisk wildcard characters are used ‘to allow for inflectional variants and to write more powerful templates with wider coverage (Piao et a. 2003)]. As a result, the MWE lexicon covers many more MWEs than the number of individual entries. Here are some **samples of MWE entries**:

spin_NN1 dryer*_NN*	B4/O3
Child*_NN* Protection_NN1 Agency_NN*	Z3c
life_NN1 expectancy_NN1	T3/X2.6
take*_* [Np/P*/R*] for_IF* granted_*	S1.2.3+
under_II [J*/R*] pressure_NN1	E6- A1.7+
*ing_NN1 machine*_NN*	Df/O2

The MWE templates are also capable of capturing discontiguous MWEs. In the fourth and fifth sample entries, the curly brackets contain words that may be embedded within a MWE. The fourth entry allows for the possibility of a noun phrase, pronoun and/or adverb occurring within the fixed phrase “take … for granted”, while the fifth entry allows an adjective and/or adverb to occur within the set phrase “under … pressure”. The last entry carries a special category “DF”, which means that the semantic category of a MWE is determined by that of its first constituent word.’

How do the MWE entries get into the Lancaster semantic lexicon?

Firstly, one has to consider the adopted definition of a MWE. A lexical unit gets MWE status in the Lancaster semantic lexicon

> *'if it expresses a meaning which is distinct from the sum of its parts or has a meaning which is difficult to recover from the sum of its parts.' (Piao et al. 2005b:380)*

In an earlier paper (Piao et al. 2003) state that

> *‘MWEs are lexical units carrying single semantic concepts’,*

(which technically could be recovered from the sum of its parts?)

Furthermore, as these definitions did not prove unambiguous in practice (Piao et al. 2003; 2005b), Biber et al.’s definition was also included: a MWE candidate was accepted

> *‘as a “good” one if it repeatedly co-occurred in the corpus and was likely to be used by different speakers/writers.’ (Piao et al. 2005b:383)*

The construction of the lexicon involves a great deal of human intervention and decision making. Both

> *'sub-lexicons have been manually constructed by linguists. The initial version has been bootstrapped from lexical resources in the CLAWS[10] system (Leech et al. 1994). [largely derived from the BNC, (Piao et al.*

[10] CLAWS (the Constituent Likelihood Automatic Word-tagging System) is a POS tagging software, which has been continuously developed since the early 1980s at UCREL (University Centre for Computer Corpus Research on Language), a research centre at Lancaster University.

> *2005b:387)] (…) For the MWE lexicon expansion, first candidate MWEs are extracted using concordance and statistical tools, then they are filtered and classified manually before being added to the MWE sub-lexicon.' (Piao et al. 2005a:no page number).*

The lexicon has been expanded over the last 10 years (Piao et al. 2005a) using a corpus driven approach and the lexicon continuous to expand. Due to the increased coverage, however, new candidate MWEs are detected slowly. According to Paul Rayson (e-mail conversation, 10 June 2007) the lexicon currently contains 18,922 MWE templates, which is 190 more templates than reported in 2005 (Piao et al. 2005a: 18,732 templates).

MWE extraction with WMatrix2

In order to extract WMatrix MWEs from a corpus the corpus has to undergo POS tagging and semantic tagging, so that word groups of the actual corpus text can be matched against the MWE templates from the lexicon. This is possible via the WMatrix2 interface[11], which uses the CLAWS tagger for POS annotation. In a second step 'USAS assigns a set of semantic tags to each item in the running text and then attempts to disambiguate the tags in order to choose the most likely candidate in each context' (Piao et al. 2003, see there and also Piao 2005b for a more detailed description of the process). The disambiguation is carried out according to seven different methods (ibid.), for instance in terms of MWEs, priority is given to semantic MWEs over single words. Furthermore, overlapping MWE templates in a sentence are dealt with six heuristics, starting off with that longer templates are preferred over shorter templates. The overall tone of the remaining heuristics is to prefer the one which starts earlier in the sentence and which uses fewer wildcards (ibid.).

[11] see http://www.comp.lancs.ac.uk/ucrel/wmatrix/ for access

Suitability of WMatrix2 MWE extraction for NICLE(s) and ENSIC

One of the great advantage over statistical approach such as WordSmith Tools' clusters is that it can find low frequent MWEs which ‘form major parts of MWEs in most corpora’ (Piao et al. 2003), Piao et al report that 68% of the accepted MWEs extracted by USAS/WMatrix occur only once or twice in the test data. Also, some templates allow for flexible/discontinuous MWUs, as could be seen in the examples of the MWE templates.

Furthermore, candidates, such as the ‘I I I’ cluster found by WordSmith Tools, are excluded, however, by relying on a more or less predefined set of MWE in the lexicon WMatrix might miss out on MWUs, especially more idiosyncratic but frequent units produced by learners. On the same note, the lexicon is not 'complete' and especially considering that there is no clear cut definition it probably never will be complete. At a first glance it does not seem too big a problem as the coverage is reported as good. In practice however, the program missed the sequence 'I don't know' for example, which is recognised as a very frequent MWU by other researchers (see e.g. Biber et al.). Thus the limitation relating to the lexicon knowledge and the dependency on the training data has to be considered in the evaluation of the results.

Along the same lines is the following observation: The same MWU/cluster can be found by both methods, however, that does not mean that they find exactly the same candidates. For example, both methods extracted 'you know' as a MWE/cluster, but WordSmith Tools found many more instances than WMatrix2. This is most likely due to the additional analysis of the semantic context by WMatrix2.

Some problems were encountered when experimenting with a test corpus (METER sub-corpus, 250,000 words; see Piao et al. 2003; 2005 for the tests). USAS extracted 4,195 MWE candidates and – acknowledging the absence of a ‘clear cut’ definition of MWEs and the problem of manually checking and disagreeing as to whether is a candidate a ‘good’ MWE or not – 3,792 MWE were considered good MWEs (90.93%). Compared with a smaller set of test data (14,711 words) which was annotated manually for MWEs, WMatrix extracted 595 good MWEs from the test data whereas manually 1,511 were found.

The absence of a clear definition of MWEs seems to be a major problem in compiling the lexicon. But nevertheless, the vast majority of MWEs which are being extracted with WMatrix seem to be valid by human intuition. The more serious limitation is that some MWEs are not detected as they are not in the lexical database.