A Large Semantic Lexicon for Corpus Annotation

Scott S.L. Piao[1], Dawn Archer[2], Olga Mudraya[3], Paul Rayson[1], Roger Garside[1], Tony McEnery[3], Andrew Wilson[3]

Computing Dept., Lancaster University, UK[1]
Department of Humanities, University of Central Lancashire, UK[2]
Dept. of Linguistics and English Language, Lancaster University, UK[3]
{s.piao; o.moudraia; p.rayson; t.mcenery; r.garside; a.wilson}@lancaster.ac.uk, dearcher@uclan.ac.uk

Abstract

Semantic lexical resources play an important part in both corpus linguistics and NLP. Over the past 14 years, a large semantic lexical resource has been built at Lancaster University. Different from other major semantic lexicons in existence, such as WordNet, EuroWordNet and HowNet, etc., in which lexemes are clustered and linked via the relationship between word/MWE senses or definitions of meaning, the Lancaster semantic lexicon employs a semantic field taxonomy and maps words and multiword expression (MWE) templates to their potential semantic categories, which are disambiguated according to their context in use by a semantic tagger called USAS (UCREL semantic analysis system). The lexicon is classified with a set of broadly defined semantic field categories, which are organised in a thesaurus-like structure. The Lancaster semantic taxonomy provides a conception of the world that is as general as possible as opposed to a semantic network for some specific domains. This paper describes the Lancaster semantic lexicon both in terms of its semantic field taxonomy, lexical distribution across the semantic categories and lexeme/tag type ratio. As will be shown, the Lancaster semantic lexicon is a unique and valuable lexical resource that offers a large-scale general-purpose semantically structured lexicon resource, which can have various applications in corpus linguistics and NLP.

1. Introduction

Lexical resources play an important part in both corpus linguistics and NLP (natural language processing). Over the past years, large semantic lexicons such as WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), HowNet (http://www.keenage.com), etc. have been built and applied to various tasks. During the same period of time, another large semantic lexical resource has been in construction at Lancaster University, as a knowledge base of an English semantic tagger, named USAS (Rayson et al. 2004)[1]. So far, the Lancaster semantic lexicon has grown into a large lexical resource, which contains over 45,800 single word entries and over 18,700 multi-word expression template entries. Employing a semantic field analysis scheme, this lexicon links English lexemes and multiword expressions to their potential semantic categories, which are disambiguated according to their context in use.

[1] The semantic lexicon and the USAS tagger are accessible for academic research as part of the Wmatrix tool, for more details see http://www.comp.lancs.ac.uk/ucrel/wmatrix/

The Lancaster semantic lexicon classifies lexemes under a set of broadly defined semantic field categories such as “food and farming”, “Life and living things”, etc., which are organised, in turn, in a thesaurus-like structure (cf. WordNet and EuroWordNet, in which lexemes are clustered and linked via the relationship between word/MWE senses or definitions of meaning). While word sense indisputably provides the substantial information for linking and organising words, the semantic field (or lexical field) identifies “named area[s] of meaning in which lexemes interrelate and define each other in specific ways” (Crystal 1995) and, as such, has long been used as a framework for structuring lexemes: see, for example, “Roget’s Thesaurus of English words and phrases” (Roget 1852), the Longman Dictionary of Scientific Usage (Godman and Payne 1979), Tom McArthur's Longman Lexicon of Contemporary English (1981), the Longman Language Activator (Summers 1993), and more recently the Cambridge Advanced Learner's Dictionary SMART thesaurus encoding (2003).

Yet, the Lancaster scheme is different from many semantic field taxonomies in use today, not least because it is conceptually rather than content driven. That is, it provides a conception of the world that is as general as possible as opposed to a semantic network for specific domains. Indeed, the lexical items subsumed within the various semantic field classifications have largely been derived from large corpora, as a means of ensuring that the lexicon better reflects real-world language usage. That said, the classification of the taxonomy is such that the automatic extraction of terminologies for various domains (*Health* and *Disease*, *Plants*, etc) is also possible. Although the terminology for some of these domains is rather limited at present, we are involved in projects that will ensure a more extensive coverage of particular domains and subjects in the near future.

In the following sections, we describe the Lancaster semantic lexicon both in terms of the structure of its semantic field taxonomy, lexical distribution across the semantic categories, and lexeme/tag type ratio. As will be shown, the Lancaster semantic lexicon is a unique and valuable lexical resource that offers a large-scale general-purpose lexicon structured according to semantic field classifications. Very importantly, it also enables the (semi-) automatic semantic field analysis of large corpus data and, as such, has various applications in the areas of corpus linguistics and NLP.

2. Lancaster Semantic Field Annotation Scheme

The Lancaster semantic field analysis scheme was initially derived from McArthur's Longman Lexicon of Contemporary English (1981), which extracts approximately 15,000 words relating to “the central vocabulary of the English language” (McArthur, 1981: Preface) from the 1978 edition of the Longman Dictionary of Contemporary English and arranges them into 14 semantic fields (or major codes). These fields are further divided into a total of 127 group codes and 2,441 set codes. For example, the “Travel and Visiting” field has sub-groups of words classified as “visiting”, “meeting people and things”, “visiting and inviting”, etc (McArthur 1981; Jackson and Amvela 2000: 112).

The Lancaster semantic field taxonomy initially utilised the same basic format, but modified and expanded the semantic divisions (see Archer *et. al*. 2004 for a comparison of McArthur's scheme and the USAS scheme). The Lancaster semantic taxonomy has since undergone further revision in the light of practical tagging problems met in the course of ongoing research. Currently it contains 21 major semantic fields that expand into 232 sub-categories. Table 1 below shows the major fields, their definitions and letter denotations. These letters form the basis of the semantic tagset used by the USAS semantic tagger.

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and the public domain	**H** Architecture, buildings, houses and the home	**I** Money and commerce in industry
K Entertainment, sports and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Linguistic actions, states and processes	**S** Social actions, states and processes
T Time	**W** The world and our environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammatical words			

Table 1: Lancaster 21 major semantic fields

The lexemes within each of the above semantic fields are further divided into areas of meaning which reflect synonym-antonym, general-specific or meronymy/holonymy relationship. For example, the {F: Food and Farming} field in Table 1 is further decomposed into four smaller meaning areas, as shown below:

F: FOOD & FARMING

F1 *Food:* Terms relating to food and food preparation
Examples: afters, bacon, banana, befores, breakfast, butter, casseroled, cereal, chilli, cook, afternoon tea, apple sauce, after dinner mint, canteen meal, chewing gum, cooking facilities, dairy product

F2 *Drinks:* Terms relating to drinks and drinking
Examples: *alcoholic, ale, beer, beverage, boozing, cola, coffee, cuppa, inebriated (++), temperance (-),*
apple juice, cherry coke, cup of coffee, drinking chocolate, glass of wine, hit the bottle, liqueur coffee, mineral water, on the wagon (-), pub crawl, Tia Maria, tonic water

F3 *Cigarettes and drugs:* Terms relating to cigarettes and (non-medicinal) drugs, including the effects of
Examples: *cannabis, cigar, detox, drugged, e-ing, LSD, non-addictive, OD,tobacco, pipe, heroin,*

cocktail cigarette, drug addiction, glue sniffing, hard drug, non smoking, passive smoking, take a puff

F4 *Farming & Horticulture* Terms relating to agriculture and horticulture
Examples: *agricultural, beehive, compost, dairy, farming, forestry, gardening, harvest, Bee keeping, estate management, free range, grounds maintenance, landscape gardening, stud farm*

As shown in this sample, *bacon*, *cereal*, *beer*, *coffee*, *cigar*, *tobacco*, *beehive*, *heroin* all fall within the semantic field of {Food and Farming}, but they can be classified into four different sub-areas of meanings: *Food*, *Drinks*, *Cigarettes and drugs*, *Food and Farming and Horticulture*, as illustrated by Fig. 1.

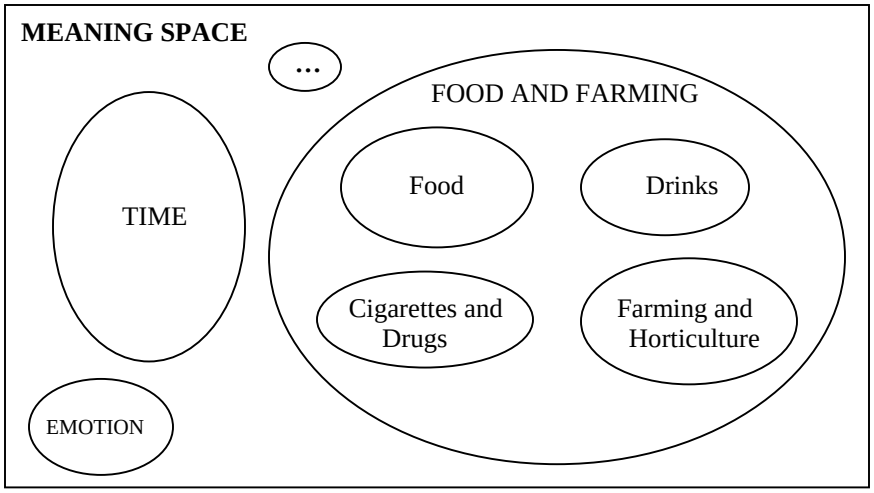

Fig. 1: Illustration of semantic relationship between lexemes

As well as providing a means of automatically extracting terminology for specific domains such as *Food*, *Entertainment*, *Sports and Games*, etc. or more specific sub-domains such as *Drink*, *Furniture and Household fittings* etc., the approach we adopt makes it possible to automatically cluster words in running texts into groups of different semantic fields/domains (general and specific). As the lexicon continues to expand, it will cover wider range of terminologies.

In theory, it is possible to include as many layers of sub-division of meaning until no further sub-classification is possible. However, excessively complex semantic field analysis schemes may cause problems for practical lexical analysis. We believe that it is better to maintain a relatively low level of granularity. Accordingly, we constrain the depth of our semantic hierarchical structure to a maximum of three layers.[2] Even so, a certain level of ambiguity and overlapping of the semantic categories remains unavoidable, not least because "English vocabulary is not made up of a number of discrete lexical fields in which each lexeme finds its appropriate place". Put simply, language cannot always "be analysed into well defined and watertight categories" (Jackson and Amvela 2000: 15). In many of these cases, we use portmanteau tags (i.e.

[2] For the full USAS semantic field taxonomy, see http://www.comp.lancs.ac.uk/ucrel/usas/

assign particular items to two – at the most three – semantic categories simultaneously) or, in the case of polysemous words, assign lexemes to the various semantic fields that most accurately capture their senses. When lexemes are assigned multiple senses, the 'correct' sense is disambiguated later in the tagging process. The tagging and disambiguation process is not the focus of this paper. It is described and evaluated in Rayson et al (2004).

3. Semantic Tagset and Lexical Entry

The semantic field information in the Lancaster lexicon is encoded using a set of semantic tags. An encoding convention has been developed to facilitate automatic processing and human comprehension.

As mentioned previously, the twenty-one uppercase letters denoting the top semantic fields form the basis of the semantic tagset. Digits are used to indicate the sub-divisions of the top semantic fields, e.g. *T1.1.1* denotes a subcategory: {*Time -> General -> Past*}. The points between the digits indicate the number of layers of sub-division, e.g. two points of the tag *T1.1.1* indicate that this category is located at the third layer of sub-division. Different granularities of semantic analysis are applied to different semantic fields. For example, while the semantic field {*A: General and Abstract Terms*} branches down into 48 sub-categories spanning over three layers of sub-divisions, the field {*L: Life and Living Things*} has only three sub-categories with only one layer of sub-divisions: *L1* (life and living things), *L2* (living creatures generally) and *L3* (plants). All together, 232 tag types are used to denote the sub-categories of the semantic field taxonomy.

In addition, a set of codes is used to denote minor semantic variance between lexemes. These codes provide a flexible way of annotating a greater diversity of semantic information than the basic 232 semantic categories. For example, an antonymous relation is indicated by +/- markers on tags; comparatives and superlatives receive double and triple +/- markers respectively. Certain lexemes show a clear double (or in some cases, triple) membership of categories. In such cases, a slash is used to combine the double/triple membership categories into what we call a portmanteau category named after similar combinations at the part-of-speech level (Leech et al, 2004) (e.g. anti-royal = E2-/S7.1+, accountant = I2.1/S2mf, bunker = G3/H1 K5.1/W3). Lower case '*i*' indicates a semantic idiom or MWE such as a phrasal verb, compound noun, etc; lower case '*f*', 'm' and '*n*' indicate 'female', 'male' and 'neuter' respectively. A rare semantic category of a word is marked with codes '*%*' or '@'.

The Lancaster semantic lexicon consists of two main parts: a single word sub-lexicon and a multi-word expression (MWE) sub-lexicon. In the single word sub-lexicon, each entry maps a word, together with its POS category, to its potential semantic categories. For example, as shown in Fig. 2, the word "iron" is mapped to the category of {*S1.2.5+: Toughness*; *Strong/Weak*} when it is used as an adjective, to the categories 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.

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

Fig. 2: Sample of single word entries

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 MWE 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. As a result, the MWE lexicon covers many more MWEs than the number of individual entries. Fig. 3 below shows some sample 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

Fig. 3: Sample of MWE entries

As shown in Fig. 3, MWE templates are also capable of capturing discontinuous 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.

For those entries to which multiple candidate semantic categories apply, the categories are arranged in a sorted sequence according to the likelihood and frequency of their application. For each lexeme, usually one or more semantic categories constitute the core or central meaning area while the others form marginal meaning area[s]. In practice, the most likely or common semantic category is put at the front of the candidate list, and the least common one is put at the end of the candidate list. Such a sorting is based on both human expert judgement and empirical statistical information extracted from corpora. Although the relative importance of the semantic categories for a given lexeme may vary in different domains, we assume that the sorting sequence used for the Lancaster semantic lexicon by and large reflects the general situation in ordinary English language usage.

With regards to the granularity of the semantic tagset, the existing taxonomy allows the lexicon to embody distinctions which are more coarse-grained than some word sense distinctions listed in standard printed dictionaries. It should be noted that fine-grained semantic distinctions, e.g. between 'bank' as a financial institution and 'bank' as a branch of a financial institution, may not be required for many NLP applications. The effect of changing granularity in sense inventories on the accuracy of word sense disambiguation can be seen in Tufis and Ion (2005).

The single-word and MWE sub-lexicons have been manually constructed by linguists. This initial versions were bootstrapped from lexical resources in the CLAWS system (Leech et al 1994). A corpus-driven approach has been adopted to the expansion of the single-word lexicon during applications of the USAS tagger to a wide range of spoken and written corpora. Moreover, during the semantic classification of unknown words into the USAS taxonomy, the linguists have been assisted by a number of knowledge sources and tools, such as concordance lines from representative corpora (such as the BNC) and printed and electronic editions of large dictionaries (such as the Collins English Dictionary). 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 (forthcoming) describes this corpus-driven approach to the detection of new candidate MWEs. Whilst we still tend to find a small percentage of unknown words when applying the tagger to new texts, its lexical coverage has been significantly improved through continual expansion of the lexicon over the past ten years. Further details of coverage of the lexicon are described in Piao et al. (2004).

As described thus far, the Lancaster semantic lexicon employs a rather comprehensive and flexible annotation scheme for inputting and retrieving lexical semantic information. In particular, such an annotation mechanism supports the automatic semantic tagging and analysis of text[s] at the semantic field level. Semantic analysis employing this tagger has been used for a variety of applications, e.g. content analysis (Thomas and Wilson, 1996) and information extraction from software engineering documents (Sawyer et al, 2002).

4. Lexical distribution across semantic categories

As lexemes within the Lancaster lexicon are classified by the semantic fields to which they belong, they can be linked – via their semantic tag[s] - to other lexemes with which they share a sense relationship. As such, word distribution across semantic fields and the lexeme/tag type ratio, that is, the balance of the lexicon and ambiguity level of the annotation are important issues for our lexicon construction and application.

We can examine the word distribution under each of the semantic fields by collecting the number of entries for each of the top 21 semantic fields. In the entries where multiple candidate semantic tag types occur, we assume that the first tag is the most representative semantic category. Table 2 shows the overall distributional structure of both the single word (second column) and MWE (third column) sub-lexicons while Fig. 4 illustrates the

distribution with a bar chart. In the table, the "DF" category applies only to MWE entries. As table2 highlights, the Lancaster lexicon currently contains 45,870 single word entries and 18,732 MWE entries.

Top semantic fields	**Single word entries**	**MWE entries**
A: general and abstract terms	6,082	2,160
B: the body and the individual	2,487	1,141
C: arts and crafts	258	110
E: emotion	1,803	582
F: food and farming	1,305	652
G: government and the public domain	1,578	781
H: architecture, buildings, houses and the home	801	430
I: money and commerce in industry	1,408	891
K: entertainment, sports and games	959	815
L: life and living things	1,024	222
M: movement, location, travel and transport	2,558	1,552
N: numbers and measurement	1,889	714
O: substances, materials, objects and equipment	3,348	600
P: education	362	316
Q: linguistic actions, states and processes	2,425	784
S: social actions, states and processes	4,284	1,559
T: time	1,157	818
W: the world and our environment	481	97
X: psychological actions, states and processes	2,529	1,036
Y: science and technology	303	255
Z: names and grammatical words	8,829	3,137
DF: MWE's first word determines its category	0	78
Total	45,870	18,732
Average (*DF excluded*)	2,184.3	888.3

Table 2: Lexeme distribution across the top semantic fields

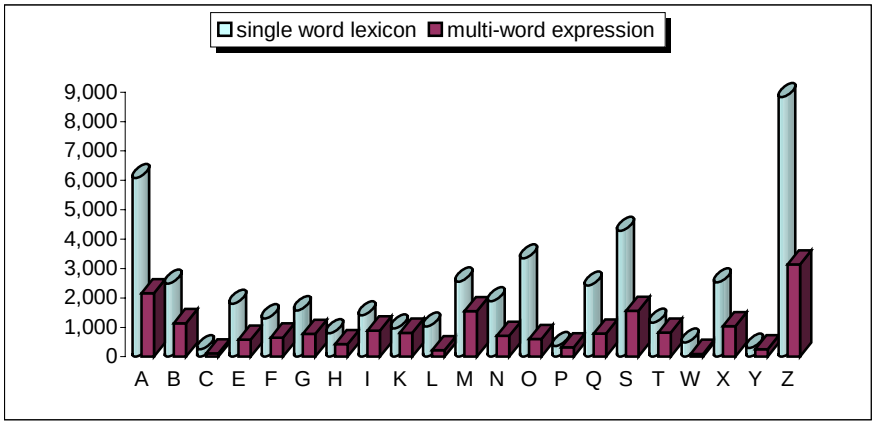

Fig. 4: Distribution chart of lexicon entries across 21 semantic categories

With reference to the single word sub-lexicon, the semantic field of *Names and Grammatical Terms*, denoted by letter *Z*, forms the largest single group, covering 8,829 entries and constituting 19.25% of the total single word entries. The category of *Arts and*

Crafts Terms, denoted by the letter *C*, forms the smallest group, containing only 258 entries. When we sort the entries in descending order by their sizes, the top ten larger semantic fields (*Z*, *A*, *S*, *O*, *M*, *X*, *B*, *Q*, *N* and *E*) contain the major part of the entries (i.e. 36,234 entries or 79.00% of the total). The remaining seven smaller fields contain 9,636 entries, that is, 21.00%.

While we cannot expect an absolutely balanced lexeme distribution for semantic categories, the abnormal size of the *Z* category is obviously not desirable. We should point out, however, that the main reason for such an unbalanced size of this category is that it covers all name entities, which, potentially, can be further divided. In fact, a number of algorithms and systems have been developed with the specific purpose of detecting and classifying named entities (e.g. Maynard et al 2003). Another reason for the abnormal size of the Z category is that it is also captures grammatical words (e.g. pronouns, negative terms, conditionals, etc.).

The Z category also forms the largest semantic field in respect of the MWE sub-lexicon. Indeed, it contains 3,137 entries (16.75% of the total). The *World and Environment* category, denoted by letter *W*, forms the smallest semantic field, containing only 97 entries (0.52%). Again, the top ten larger categories (*Z*, *A*, *S*, *M*, *B*, *X*, *I*, *T*, *K* and *Q*) contain 13,893 entries (74.17% of the total). It should be noted that many MWE templates represent multiple MWEs using wildcards, and therefore the actual number of MWEs covered can be much larger than the number of MWE entries.

In fact, the top twenty-one semantic fields denoted by twenty-one letters only provide a general framework for the semantic scheme. As we explained in section 2.2, the letters, digit numbers indicating subdivisions and some auxiliary codes such as '+' and '-' can be combined to form numerous tag types to depict fine-grained semantic categories and minor variance of semantic features. When we examined the semantic tags in the lexicon (again, only the first tag was considered where multiple candidate tags occur), we found that 2,999 and 2,763 tag types respectively occur in the single word and MWE sub-lexicons. Of particular interest to us is the number of entries covered by each tag type, or the lexeme/tag type ratio. Such a ratio can be used as an indicator of the ambiguity level of the semantic field analysis. A higher ratio would indicate a higher multiplicative ambiguity, and vice versa. In this regard, then, Table 3 shows the frequency distribution of tag types in relation to the number of lexicon entries they cover. The left-hand columns relate to the single word sub-lexicon and the right-hand columns to the MWE sub-lexicon.

As shown in the left half of the table (i.e. the sections relating to the single word sub-lexicon), only 99 tag types (3.30%) cover more than 100 single word entries each. In contrast, 2,159 (71.99%) tag types are used as the first (i.e. most representative) tag in only three or fewer entries. Moreover, 1,482 tag types (49.42%) occur as the main semantic tags just once. When we consider the total single word entries (= 45,870), the average lexeme/tag type ratio is 45,870 ÷ 2,999 = 15.30. In fact, 103 sub-categories under the *Z* semantic field alone cover 8,829 entries. More specifically, tags Z2 (Geographical Names), Z3c (Other Proper Names) and Z1mf (Personal Names) cover 2,880, 1,566 and

1,368 entries respectively. If we ignore Z-initial categories, the lexeme/tag type ratio drops to (45,870 – 8829) ÷ (2,999 – 103) = 12.79. As 71.99% of the tag types have lexeme/tag type ratios ranging between three and one, we assume the single word lexicon has a rather low ambiguity level.

Single word sub-lexicon		MWE sub-lexicon	
Number of entries covered	Number of tag types for each range of coverage	Number of entries covered	Number of tag types for each range of coverage
>= 101	99	>= 101	21
81 – 100	19	81 – 100	11
61 – 80	34	61 – 80	26
41 – 60	56	41 – 60	40
21 – 40	94	21 – 40	87
11 – 20	139	11 – 20	85
4 – 10	399	4 – 20	254
3	223	3	166
2	454	2	457
1	1,482	1	1,616
Total types	2,999	Total types	2,763

Table 3: Lexicon entries vs. tag type distribution

In regard to the MWE sub-lexicon (see right half of Table 3), a total of 2,763 tag types were found. As a result, the average lexeme/tag type ratio is 18,732 ÷ 2,763 = 6.78, which is much lower than the single word sub-lexicon. This lower ratio is not surprising, as MWEs contain multiple constituent words, and in consequence are less ambiguous than single word items. In fact, few tag types cover a large number of entries. For example, Z3c, Z2 and M1 (Moving, Coming and Going) contain 1,328, 674 and 477 entries respectively. When we ignore 63 *Z*-initial tag types which cover a total of 3,137 entries, the lexeme/tag type ratio drops to (18,732 - 3,137) ÷ (2,763 - 63) = 5.78. As with the single word sub-lexicon, the majority of MWE tag types have narrow entry coverage. Indeed, 2,239 tag types, or 81.04% of the total, occur as the main semantic category in three or fewer entries, that is, they have a lexeme/tag type ratio equal to or lower than three.

As Table 2 and 3 reveal, the lexicon contains a fairly large number of words and MWEs for many of the semantic domains, e.g. 2,299 terms for {I: Money and Commerce in Industry} and 2,603 terms for {N: numbers and measurement}, with an average of 2,184. These lexemes can be extracted at the general or specific levels, making terminology extraction more viable. Given the fact that many terminological terms are multiword units (e.g.. “TCL screen”), the MWE templates also provide a very useful means of extracting multi-word terms. Nevertheless, the current lexicon is not well balanced across domains for this purpose as yet. The domain of {C: Arts and Crafts}, for example, has only 368 items. This problem will be alleviated as the lexicon is expanded further.

As we have shown, the Lancaster lexicon covers a wide range of semantic domains. Although the distribution of the lexemes across the semantic categories is not well

balanced yet, each of the twenty-one top domains contains an essential part of its core terms. Also, we have seen a reasonably low lexeme/tag type ratio, even lower for the MWE sub-lexicon, demonstrating that this lexicon provides a practically useful resource for semantic disambiguation. The main problem we have found relates to the *Z*-initial categories, which contain an un-proportionally large number of lexemes. One possible solution can be to further classify these particular lexemes into more fine-grained semantic categories.

5. Conclusion

In this paper, we have presented the Lancaster semantic lexicon in terms of its semantic field taxonomy, lexical distribution across the semantic categories and lexeme/tag type ratio. We have shown that the Lancaster semantic lexicon is a unique lexical resource. Indeed, adopting the semantic field as the organising principle of lexical structure, it has distinct features from other major semantic lexicons in use today. In particular, its design facilitates the automatic word classification and extraction of terminology for a number of semantic domains. It is also able to capture many MWEs via its templates.

Our research on the semantic lexicon is continuing. We have completed the first version of a Finnish semantic tagger by porting the Lancaster semantic taxonomy (Löfberg et al 2005), and we are currently reviewing the contents of certain semantic categories to improve the consistency of the English lexicon. Furthermore, we are developing a Russian semantic tagger which will form one component of a translator-assistant tool. We envisage that the Lancaster semantic lexical resource will have various additional applications in the areas of corpus linguistics and NLP.

Acknowledgement

The work presented in this paper has been carried out as part of the Benedict Project funded by the European Community (ref. IST-2001-34237) and the Assist Project funded by EPSRC in the UK (EP/C004574/1).

References

Archer, D., Rayson, P., Piao, S., McEnery, T. (2004) Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies, in *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004)*, Lorient, France Université de Bretagne Sud. Volume III, 817-827.

Cambridge Advanced Learner's Dictionary (2003). Cambridge University Press.

Crystal, D. (1995) *The Cambridge Encyclopaedia of the English Language*. Cambridge University Press.

Fellbaum, C. (ed). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Goodman, A. and Payne, E. (1979) *Longman Dictionary of Scientific Usage*. Hong Kong, Longman.

Jackson, H. and E. Zé Amvela (2000) *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. The Cromwell Press, Trowbridge, UK.

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan*, 622-628.

Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P, Nykanen, A., and Varantola, K. (2005) A semantic tagger for the Finnish language. In *proceedings of Corpus Linguistics 2005, Birmingham, July 2005*.

McArthur, T. (1981). *Longman Lexicon of Contemporary English*. Longman, London.

Maynard, D., Tablan, V., and Cunningham, H. (2003). NE recognition without training data on a language you don't speak. *ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models*, Sapporo, Japan.

Piao, S., Rayson, P., Archer, D., McEnery, T. (2004). Evaluating Lexical Resources for A Semantic Tagger. In *proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004), May 2004, Lisbon, Portugal*, Volume II, 499-502.

Piao, S., Rayson, P., Archer, D., McEnery, T. (forthcoming) Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language*, Elsevier.

Rayson, P., D. Archer, S.L. Piao, T. McEnery (2004) The UCREL Semantic Analysis System, in *Proceedings of the LREC-04 Workshop, Beyond Named Entity Recognition Semantic labelling for NLP tasks*, Lisbon, Portugal. May 2004, 7-12.

Roget, P. M. (1852) *Roget's Thesaurus of English words and phrases*. Longman.

Sawyer, P., Rayson, P., and Garside, R. (2002) REVERE: support for requirements synthesis from documents. *Information Systems Frontiers Journal. Volume 4, Issue 3*, Kluwer, Netherlands, 343 - 353.

Summers, D. (ed.) (1993) *Longman Language Activator*. Longman.

Thomas, J., and Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas and M. Short (eds) *Using corpora for language research*. Longman, London, 92-109.

Tufis, D. and Ion, R. (2005). Evaluating the word sense disambiguation accuracy with three different sense inventories. In B. Sharp (ed.) *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science (NLUCS2005)*, 118-127.

Vossen, P. (1998) Introduction to EuroWordNet, in Nancy Ide, Daniel Greenstein, Piek Vossen (eds.) Special Issue on EuroWordNet. *Computers and the Humanities* (32: 2-3), 73-89.