Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies

Dawn Archer^A, Paul Rayson^B, Scott Piao^A and Tony McEnery^A
Dept. of Linguistics and Modern English Language^A and Department of Computing^B
Lancaster University,
 Bailrigg Campus,
 Lancaster, UK
d.archer/p.rayson/s.piao/t.mcenery{@lancaster.ac.uk}

Abstract
Annotation schemes for semantic field analysis use abstract concepts to classify words and phrases in a given text. The use of such schemes within lexicography is increasing. Indeed, our own UCREL semantic annotation system (USAS) is to form part of a web-based 'intelligent' dictionary (Herpio 2002). As USAS was originally designed to enable automatic content analysis (Wilson and Rayson 1993), we have been assessing its usefulness in a lexicographical setting, and also comparing its taxonomy with schemes developed by lexicographers. This paper initially reports the comparisons we have undertaken with two dictionary taxonomies: the first was designed by Tom McArthur for use in the Longman Lexicon of Contemporary English, and the second by Collins Dictionaries for use in their Collins English Dictionary. We then assess the feasibility of mapping USAS to the CED tagset, before reporting our intentions to also map to WordNet (a reasonably comprehensive machine-useable database of the meanings of English words) via WordNet Domains (which augments WordNet 1.6 with 200+ domains). We argue that this type of research can provide a practical guide for tagset mapping and, by so doing, bring lexicographers one-step closer to using the semantic field as the organising principle for their general-purpose dictionaries.

1. Introduction
Semantic annotation – semantic field analysis, in particular - is increasingly being used within lexicography, as a means of distinguishing between the lexicographic senses of the same word. The reason, as Jackson and Zé Amvela (2000: 112) highlight, is that a ‘semantic field arrangement brings together words that share the same semantic space’, and thus provides ‘a record of the vocabulary resources available for an area of meaning’. This, in turn, enables ‘a user of the language, whether a foreign learner or a native speaker, to appreciate often elusive meaning differences between words’. Yet, as Jackson and Zé Amvela also highlight (2000: 113), there is as yet no general-purpose dictionary that uses the semantic field as its organizing principle (but see section 6). Indeed, lexicons using the semantic field principle tend to be based on religious texts and/or thesaurus-like in nature (e.g. Louw-Nida and Hallig-Wartburg-Wilson).1 It’s worth noting that, many of these semantic category systems agree, to a greater or lesser extent, on the basic major categories that they contain. However their structure and granularity are very different (cf. Wilson and Thomas 1997: 57). By way of illustration, the Louw-Nida model utilises 93 general categories at the top level, 71 of which contain one additional sub-category (see http://www.comp.lancs.ac.uk/ucrel/usas/louw-nida.htm for a full list of the general categories). In contrast, the Hallig-Wartburg-Wilson model has only three general

817
categories, the ‘universe’, ‘man’ and ‘man and the environment’. However, each general category contains four or five levels of sub-categories, many of which contain fine-grained distinctions (see http://www.comp.lancs.ac.uk/ucrel/-usas/www.htm for the full list).

In this paper, we will be concentrating on schemes that are more general in their approach than the above, by which we mean, they (purport to) presuppose a thorough conceptual/semantic analysis of their potential members and the relations between them. We do so, initially, to determine whether general structures differ greatly from more domain-specific ones. In pursuit of this, section 2 describes the taxonomy developed by Tom McArthur (1981) for use in the Longman Lexicon of Contemporary English (henceforth LLOCE), and section 3 describes the taxonomy developed by Collins for the Collins English Dictionary (henceforth CED). Section 4 then describes the UCREL semantic analysis system (USAS), the initial tagset of which was loosely based on LLOCE, but has since been revised in light of practical tagging problems met in the course of ongoing research. We also discuss ongoing work on the Benedict project to determine the possibility of mapping the USAS system to the subject field codes used in the CED (see section 5), before assessing the possibility of mapping to other systems, in particular, WordNet and WordNet Domains (see section 6). Our motivation for engaging in comparative analysis of this nature is three-fold. Firstly, we want to assess the usefulness of USAS in a lexicographical setting. Secondly, we see such work as a way of reviewing (and improving) the USAS system. Thirdly, we believe that comparative analyses of this type can bring lexicographers one-step closer to using the semantic field as the organising principle for their general-purpose dictionaries, by providing a guide for practical tagset mapping.

2. The Longman Lexicon of Contemporary English

LLOCE is a relatively small thesaurus, containing some 15,000 words, so why are we including the scheme as an example of a general taxonomy? We do so for two reasons. Firstly, because the design purports to be ‘of a pragmatic, everyday nature’ (Preface, p. vi), and therefore appears to presuppose a thorough conceptual/semantic analysis of its potential members and the relations between them. By this we mean that it not only attempts to determine the different senses for every word relevant to a text or texts under consideration, but also aims to capture all potentially relevant words in some way (Ide and Véronis 1998: 3). Secondly, as previously explained, the USAS taxonomy was originally based on LLOCE (see section 4.1).

Like the domain-specific models (above), LLOCE is hierarchical in structure, having fourteen major codes, 127 group codes and 2,441 set codes (the set codes are classified according to part-of-speech membership). Figure 1 provides a general idea of the semantic areas covered by LLOCE’s major codes:
If we compare LLOCE to the Louw-Nida and Hallig-Wartburg-Wilson models mentioned above, we find a similar pattern to that found when comparing the domain-specific models to each other. There are some obvious structural differences, but there are also obvious similarities in terms of content: All three models account for the same types of semantic area (i.e. man’s existence in the universe, and all that that entails; food, work, rest, reproduction, verbal/artistic/intellectual expression, etc.). Even LLOCE’s ‘entertainment, sports and games’ domain has observable overlaps with Hallig-Wartburg-Wilson’s ‘Physical Activity’ sub-category and Louw-Nida’s ‘Contests and play’ and ‘Festivals’.

One approach we considered taking in this paper was to assume that some domains must therefore be universal, and concentrate our energies on finding and investigating them alone. But we have come to believe that we can gain much by also exploring differences between taxonomies. We might, for example, concentrate on what semantic areas particular taxonomies omit or background. By way of illustration, several of Louw-Nida’s categories – including ‘Contests and play’, ‘Festivals’, ‘Agriculture’ and ‘Animal husbandry/fishing’ - are not sub-classified, suggesting that they were used very little and/or that the data was such that Louw and Nida (1989) did not have reason to fine-grain them further. It’s also worth noting that the domain-specific models do not provide a classification for ‘Art’ and its related concepts (unless the Louw-Nida model classifies this type of domain under ‘Artefacts’). Such findings lead us to conclude that, whilst the Louw-Nida model adequately accounts for the concepts that arise in the Greek New Testament, it may not capture the complete world-view (or mindset) of the specific people groups/cultures that it claims to represent. A possible solution to this is to leave the ontology ‘open’, by which we mean allow for new categories to be added and existing categories to be made more fine-grained as and when the need arises (see section 4.1). However, mention of cultural mindsets highlights another important issue, which we will touch upon at various points in this paper: the extent to which a semantic network can ever universally applied. In the following sections, we describe the Collins taxonomy (section 3) and our own USAS system (section 4).

3 The Collins taxonomy

Collins prefer the term ‘subject field’ to ‘semantic field’ when assigning sense domains in the tagged version of the Collins English Dictionary (CED). Nevertheless, the principle remains the same (i.e. bringing together words that share the same semantic space). Collins adopt seven major subject field codes:
These major fields are not explicitly coded in any way. Instead, the dictionary entries in the tagged version are coded according to related sub-fields. Although these underlying 'genus' and 'subject' fields enable Collins to extract sets of vocabulary relating to specific subject areas (the printed CED does not contain coding at this level), it's worth noting that (i) the 'General' domain is not sub-divided (and therefore left un-coded), (ii) codes relating to the remaining semantic groups are not applied to all words systematically (rather, words are given codes only when Collins deem them to be necessary for disambiguation purposes), and (iii) many of the words tend to be technical in nature. This means that the Collins' system captures information that is largely domain-specific, even though the taxonomy itself is conceptually based.

As part of the Benedict project, we have been exploring the extent to which the semantic coverage/sense disambiguation of the CED might be improved by mapping the USAS taxonomy to the latter's subject field codes (cf. Véronis and Ide 1990). A report of that work follows our description of the USAS system.

4 The UCREL Semantic Analysis System

The USAS system is a software package for automatic dictionary-based content analysis, and consists of:

1. CLAWS (Garside and Smith 1997), a part-of-speech tagger which assigns a part-of-speech tag to every lexical item or syntactic idiom in the text,
2. SEMTAG (Wilson and Rayson, 1993 and 1996), which assigns a semantic tag (or tags separated by slash tags, when more than one sense is appropriate) to each lexical item or multi-word unit, and
3. AUXRULE, a sub-module of SEMTAG that disambiguates the auxiliary and main verb senses of be, do and have with a high degree of accuracy on the basis of their close collocation, or lack of collocation, with specific participial forms (Thomas and Wilson 1996: 97).

The tagset of the SEMTAG element includes 21 major discourse fields, which, expand, in turn, into 232 category labels with up to three sub-divisions. Each tag is represented by a decimal notation; the major discourse field is shown by a capital letter (see Figure 3 below), the subdivisions by numerals (e.g. L2 [= 'living creatures generally']), and further subdivisions by further numerals separated off by points (e.g. S1.2.3 [= 'egoism']).
6.1 Criteria underlying the UCREL Semantic Analysis System

Although there is no such thing as an ideal semantic annotation scheme, Wilson and Thomas (1997: 55-6) suggest that a workable taxonomy should:
1. Make sense in linguistic or psycholinguistic terms.
2. Account exhaustively for the vocabulary in the corpus.
3. Be sufficiently flexible to allow for necessary emendations.
4. Operate at an appropriate level of granularity (or delicacy of detail).

As will become clear, these features have greatly influenced the design and development of USAS.

The original USAS ontology was largely based on LLOCE, as it appeared to offer the most appropriate thesaurus-type classification of word senses for dictionary-based content analysis. Consequently, both systems have the following top-level categories in common:

<table>
<thead>
<tr>
<th>USAS</th>
<th>LLOCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>General and abstract terms</td>
<td>General and abstract terms</td>
</tr>
</tbody>
</table>

Figure 3 USAS tagset top-level domains

The 232 category labels each represent a particular semantic field or ‘space’. In simple terms, they group together senses that are related by virtue of their being connected at some level of generality with the same mental concept (whether this is via a process of synonymy, antonymy, hypernymy and/or hyponymy). The tags themselves are assigned on the basis of dictionary look-up between the text and two lexical resources developed for use with the program: a lexicon of single word forms and an ‘idiom list’ of multi-word units, which presently contain 61,400+ items. However, some fixed patterns with many possible instantiations (e.g. ‘Xkm’, where ‘X’ is a number) are tagged by automatic rules (‘Xkm’ is automatically assigned to the linear measurement category). Tests have shown that SEMTAG has a 92% accuracy rate (Piao et al 2004). Disambiguation of the correct sense is helped not only by the part-of-speech categories that CLAWS assigns, and the AUXRULE module (see above), but also by the intuitive frequency-ordering of the possible semantic categories for each word/multi-word unit in the lexical resources (see Garside and Rayson 1997).
In addition, individual top-level categories within LLOCE have been transformed into separate top-level categories in USAS. These include:

<table>
<thead>
<tr>
<th>USAS</th>
<th>LLOCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts and crafts</td>
<td>Arts and crafts, science and technology, industry and education</td>
</tr>
<tr>
<td>Science and technology</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>Numbers and measurement</td>
<td>Numbers, measurement, money and commerce</td>
</tr>
<tr>
<td>Money and commerce</td>
<td></td>
</tr>
<tr>
<td>Government and the public domain</td>
<td>People and the family</td>
</tr>
<tr>
<td>Social actions, states and processes</td>
<td></td>
</tr>
<tr>
<td>Linguistic actions, states and processes</td>
<td>Thought, communication, language and grammar</td>
</tr>
<tr>
<td>Psychological actions, states and processes</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Space and time</td>
</tr>
<tr>
<td>The world and our environment</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Top-level LLOCE categories and their USAS counterparts

As USAS automatically tags every word in a text, we have also added a category – ‘Names and grammatical words’ – that captures words that are traditionally considered to be ‘empty’ of content (i.e. closed class words) and proper nouns. The revisions reflect our responses to problems met in light of tagging English texts from a variety of domains across different historical periods (Piao et al 2004), and for a variety of purposes (e.g. market research, content analysis, information extraction, keyword extraction, etc.).

The above semantic field categories are meant to provide a conception of the world that is as general as possible (cf. ontologies that are ‘content’ driven, i.e. words are classified according to the operationalisation of a theory or research hypothesis rather than on general semantic grounds). A consequence of designing a general (as opposed to domain-specific) system is that some of the fine-grained distinctions made by other taxonomies can be lost. By way of illustration, the USAS system does not have a separate ‘birds’ category, choosing to classify all living creatures together, under a ‘living creatures generally’ category, which is a sub-category of ‘L: Life and living things’ (see Figure 3).7 This particular ‘granularity’ issue is not overly problematic, as the hierarchical design of the USAS system ensures that it can be further fine-grained as and when the need arises. By way of illustration, we might sub-divide the ‘living creatures generally’ category so that it includes separate categories for ‘creatures of the land’, ‘creatures of the sea’ and ‘creatures of the air’. These sub-categories, in turn, could be further divided, so that a distinction can be made between ‘wild birds’ and ‘domestic birds’ and ‘fish’ and ‘crustacean’). That said, one has to remember to balance the
desire for highly fine-grained distinctions with the desire to be culturally relevant (birds considered to be wild by one culture may be thought of as pets by another culture).

4.2 The MAPPING component of the UCREL Semantic Analysis System

The preference of many social scientists to carry out content analysis has led to the inclusion of a module (MAPPING) that, by enabling word + sense combinations to be mapped automatically into research-specific content categories, provides a second means of overcoming the ‘granularity’ issue (see Wilson and Thomas 1997: 55). The following section highlights work undertaken by members of UCREL and Collins for the Benedict project, using this MAPPING module.

5. Mapping between the USAS tagset and the CED tagset

The USAS MAPPING module maps the top-level categories of the USAS system to the top-level categories of the CED model as shown in Figure 6 (below). There are several things to notice here, not least the differences in concept names and taxonomic structure. Differences in the latter are potentially more problematic than differences in concept names (e.g. ‘Money and Commerce’ versus ‘Business and Economics’). Indeed, the fact that the CED system contains fewer top-level categories means that many of their categories map to more than one of the USAS top-level categories. Particular USAS top-level categories (e.g. the ‘body and the individual’) also map to one or more of the CED top-level categories (e.g. ‘arts’ and ‘science and technology’). In addition, five of the USAS top-level categories cannot be directly mapped to any of CED’s top-level categories (see ‘unmatched categories’). These factors highlight an important point, namely, that semantic categorization is always a matter of the designer[s]’ personal judgement, to some degree, not least because a sense of a particular word can be (and often is) classified into two or more semantic categories. This suggests, in turn, that one-to-one mapping of the top-level hierarchies of any system is potentially unlikely.
Although the absence of one-to-one mapping complicates the mapping procedure, mapping between USAS and the CED subject field codes is still possible, as mismatches at the top-level can be sorted out by mapping USAS sub-levels to particular content tags. By way of illustration, the ‘geographical names’ sub-category of the top-level USAS category, ‘names and grammatical words’ will map to ‘Physical Geography’, a sub-division of the CED’s ‘science and technology’ category. The full USAS tagset also provides a means of capturing the different semantic areas that are presently grouped together under CED’s ‘General’ field, but left un-coded. Moreover, as the USAS software automatically links words appearing in running text to their semantic categories, those semantic areas can be isolated so that (where necessary) new content tags can be created (see section 6).

6. Semantic fields as an organising principle: the way forward?

Lexicographers are increasingly using semantic fields as a complimentary disambiguation procedure, with promising results. For example, as part of the Benedict project, Collins are involved in the development of a bilingual dictionary, and as part of this project, UCREL and Collins have been exploring the possibility of using additional dictionary entry elements for semantic tagging purposes. In particular, we’ve been assessing the feasibility of semantically tagging synonyms, definitions and collocations as a means of disambiguating sense domains (and, thus, different senses of a particular word or multi-word-unit). Although in its early stages, this work points to the possibility of using semantic fields as the organising principle for general-purpose dictionaries (see Löfberg et al 2004, this conference, for more details). However, we believe that semantic fields will only provide an
adequate organising principle if general semantic areas as well as the more technical domains are identified/differentiated within dictionaries. USAS offers an automated means of achieving this.

Our collaboration with Collins has also led to members of UCREL investigating the possibility of mapping the USAS tagset to WordNet (Felbaum 1998). We should point out that the USAS system already uses WordNet Synonyms (sets) to help disambiguate the sense of (and thus assign tags to) words not pre-classified in the USAS lexicon. However, we wanted to assess what else we might gain by mapping the two systems. Like the USAS system, WordNet offers users a reasonably comprehensive machine-useable lexical database. However, whereas the USAS system has a hierarchical, multi-tier structure, which can be further fine-grained as and when necessary (or mapped onto other content labels, as in the case of the CED tagset), the WordNet system is a set of separate networks for different parts of speech, each of which 'consists in large part of a tree structure whose root node corresponds to the general concept, and in which paths leading down from the root traverse nodes represent increasingly specific concepts' (Felbaum 1998: 56). WordNet also lacks domain terminology. As this means that the two systems share only superficial similarities, we are looking into the possibility of mapping USAS categories to specific synonym sets within WordNet. This work, in particular, should enable us to assess how well the USAS software can be used to distinguish between WordNet synonym sets in running texts, and thus fits well with Senseval, a Word Sense Disambiguation evaluation workshop (see http://www.senseval.org).

Work being undertaken to make WordNet domain specific offers interesting possibilities of our mapping the USAS tagset to WordNet in its entirety in the near future. WordNet Domains is particularly promising. Considered to be an extension of WordNet by one of its creators, WordNet Domains augments WordNet 1.6 with 200+ domain labels, including MEDICINE, ARCHITECTURE and SPORT (Magnini et al 2002). Moreover, these domain labels are organised hierarchically, like our own system. This means that mapping to WordNet becomes much easier, as we can initially map to WordNet Domains and, from there, to WordNet.

7. Acknowledgements

This paper would not have been possible without the help of members of the Benedict team at Collins Dictionaries. We are especially grateful to them for allowing us access to their taxonomy, and for the insightful comments they have provided at various stages of this paper’s production. Needless to say, remaining errors and infelicities are ours.

8. Endnotes

1 The Louw-Nida model (1989) has been used to produce the Semantic Domain Lexicon of Greek New Testament (1989) and Heidebrecht’s (1993) Lexicon of Metal Terminology in Hebrew Scriptures. The Hallig-Wartburg (1952) scheme provided the taxonomic foundation for the Conceptual Dictionary of Mycenaean Greek (Kazanskiene and Kazanskij 1986). As our addition of ‘Wilson’ to the Hallig-Wartburg scheme implies, the scheme has since been revised by Andrew Wilson, as a means of handling the social and religious make-up of the world of the gospels (Wilson
EURALEX 2004 PROCEEDINGS

1996). It has also been revised by Klaus Schmidt, as a means of capturing the social make-up of the world of mediaeval German epic (Schmidt 1988, 1993, 1994).

2 The Benedict project seeks to cater for the demands of the multilingual corporate world, by tailoring the dictionary information supply according to user specifications, and incorporating multi-layered entry structure with new information categories and links to corpus data and syntactically- and semantically-based corpus search tools in the dictionary data base. Benedict project partners are Kielikone Oy, HarperCollins Publishers Ltd, Lancaster University, Gummerus Kustannus Oy, University of Tampere, and Nokia (funded by the European Community under the 'Information Society Technologies' Programme reference number: IST-2001-34237).

3 Véronis and Ide (1990a) undertook experiments on 23 ambiguous words in six contexts (138 pairs of words), to determine how accurately the sense distinctions in the CED correctly disambiguated the words in each context. They found that the sense distinctions proved sufficiently fine-grained 71.7% of the time, but that correct sense disambiguation rose to 90% when the senses provided by the CED were mapped to the OALD (see also Ide and Véronis 1990b).

4 When a word is not in the CLAWS lexicon, CLAWS uses probabilistic Markov models of likely part-of-speech sequences and suffix heuristics. When a word is not in the SEMTAG lexicon, SEMTAG assigns an unmatched semantic tag (i.e. Z99).

5 A full list of the categories is available online at http://www.comp.lancs.ac.uk/ucrel/usas/usas-tree.htm.

6 Although we use the term “idiom list”, the latter is comprised of not only “genuine idioms”, but also phrasal verbs, multi-word proper nouns, and other multi-word units which are felt to constitute phraseological units for the purpose of semantic analysis (see Thomas and Wilson 1996: 96).

7 This loss of granularity is not true of the LLOCE model, of course, which highlights an important fact about mapping between different systems, namely, there will never be a one-to-one mapping of the different categories (see section 5).

8 Content analysis categorisations avoid this issue due to the focus on one theory or research hypothesis.

9. References


