A METHODOLOGY FOR AUTOMATIC TERM RECOGNITION

Sophia Ananiadou

Department of Computing, Manchester Metropolitan University
John Dalton Building, Chester Street, Manchester, UK, M1 5GD

1 INTRODUCTION

The topic of automatic term recognition (ATR) is of great interest especially with the growth of NLP systems, which are passing from the development stage to the application stage. The application of NLP technology involves customising systems towards specific needs, particularly in specialised domains (sublanguages) which form the main target of the technology. There is thus an urgent need for high quality, large scale collections of terminologies (with associated linguistic information) for use in NLP system dictionaries.

The existence of coherently built terminologies leads to better performance for many interesting applications: translation, technical writing (in the mother tongue or in a foreign language), multilingual (multimedial) document production, classifying, indexing, archiving and retrieving documents (monolingual and multilingual), extracting, reorganizing and reformulating knowledge represented in textual form (Heid,U. and McNaught 1991).

Given the amount of specialised texts that need to be processed in efforts to discover (potential) terms, keep track of the life-cycle of terms, etc., it is of interest to consider the design of (semi)automatic aids. A term recognition tool would be a great aid to special lexicographers. It is only an aid, however, if it incorporates linguistic and terminological knowledge such that it makes largely accurate proposals. It is with the design of such a tool that we have been particularly concerned and we report below on various aspects with which we have had success. We do not claim to have solved the term recognition problem. As we shall see, there are many different kinds of term formations each of which calls for different techniques and different knowledge: our work has concentrated on a subset of these.

2 TERMS

The main characteristic of terms is that they are monoreferential to a very great degree: a term typically refers to a specific concept in a particular subject field. Subject field knowledge is conveyed via naturally occurring **sublanguages** which demonstrate restrictions on the lexical, syntactic and semantic levels. Each sublanguage has its own set of concepts and a terminologist will structure these into a system organised according to relationships holding between concepts (generic, partitive, causal, etc.). Terms are therefore the linguistic realisation of the concepts of special communication and are organised into systems of terms which ideally reflect the associated conceptual system.

Terms differ from general language words primarily by their nature of reference. Nevertheless, it is not always an easy task to divide terms from words. The terminologist is often faced with cases where synonymy, homonymy or polysemy could be said to be playing a role. Also, the same wordform can refer to different concepts in different sublanguages, and be used differently in general language.

In elaborating a terminology, we typically employ texts and a set of tools (e.g. KWIC, inverse KWIC, etc.) which will facilitate the tasks of identifying which words, phrases, acronyms, etc., of the corpus are functioning as terms with respect to the conceptual system of the subject field and of relating them appropriately via relationships and definitions. One important aspect in terminology is the creation of new terms (term formation) in response to a new concept – this may be by, say, an engineer (monolingually) or a translator (where a target equivalent is missing).

Term formation follows certain guidelines (procedures) which vary in sophistication depending on the subject domain. Some domains show evidence of mandatory adherence to strict term formation. Others may permit wide choice within recognized limits. We view term formation in our research from an analytical viewpoint, i.e. we investigate the linguistic form of known terms in order to identify productive means of forming terms, which is necessary in order to be able to recognize new terms.

3 A METHODOLOGY FOR AUTOMATIC TERM RECOGNITION

We investigated the relevance of other disciplines to automatic term recognition, such as Information Science – especially techniques of automatic indexing. We concluded that non-linguistic based techniques (statistical and probability based ones), while providing gross means of characterizing texts and measuring the behaviour and content-bearing potential of words, are not refined enough for our purposes. In Terminology we are interested as much in wordforms occurring with high frequency as in rare ones or those of the middle frequencies. We are interested in **all** units that may be acting as terms in a collection of texts. However, we do not deny the useful role of such techniques. They have their place in that they may usefully complement other techniques.

We chose to concentrate on potential contributions of Linguistics, especially from lexical morphology, and were interested in developing methodologies for term recognition that apply theoretically motivated ideas about term formation. Theoretical Linguistics deals exclusively with general language word structure. We designed an integrated model of word and term structure based on the results of an analysis of Immunology terms in the sublanguage of Medicine (for English) and on models to be found in the literature on general language (Selkirk, 1982; Mohanan, 1986).

Medical terminology relies heavily on Greek (mainly) and Latin neoclassical elements for the creation of terms such as 'erythrocyte' and 'angioneurotic'. In the literature of theoretical Linguistics there are no satisfactory accounts of the neoclassical vocabulary and no formal motivated classification of neoclassical wordforms exists. In Terminology, most accounts of term structure remain at an unformalised descriptive level and this is particularly true for discussions of neoclassical vocabularies.

The reason for this overall lack of formal description of neoclassical elements appears to be due to their occupying a peripheral or ambiguous place in most analyses of word and term formation in English. We found this to be unsatisfactory for the following reason: it is anomalous to conceive of English word formation as being somehow separated from term formation, especially as terms constitute the majority of English words. Therefore, we strove to set up an integrated model of word and term structure which would, importantly, account adequately for the neoclassical component.

The word structure of English can be said to comprise 3 category types, i.e. Word, Root and Affix (Selkirk, 1982)[1].

However, there is great confusion in the literature as to the morphological status of Greek and Latin neoclassical forms, i.e. whether they are roots, affixes or even both. Models which describe them as affixes allow the generation of forms such as *affix+affix. Many models, including the unformalised ones of conventional dictionaries, characterise neoclassical elements vaguely as 'combining elements', which suggests some kind of extra-morphological status (or wastebasket status). Such forms thus apparently defy attempts to provide an integrated account in terms of the accepted morphological categories.

In our approach, we introduced a *fourth* category type **comb**, to help handle the neoclassical wordstock of English. This does not, in itself, resolve the problem of how to (sub)classify neoclassical elements: we will address this aspect below in detail. Firstly, though, we discuss our concomitant adoption of a *level ordered* approach to the morphological analysis of English words and terms.

Level ordering places strong constraints on the cooccurrence or order of classes of affix and hence is a powerful mechanism in helping to identify whether a wordform is well-formed or not, whether a wordform may be segmented in a particular way or not, etc. Numerous models incorporating level ordering have been proposed in morphology and morphophonology. There is debate on how many levels should be identified and what the relationships between levels are. Level ordering has its critics also. We do not enter into these debates here, however we have found, in experiments over the years, that level ordering is of great use in a computational morphology environment, as has been recently also suggested by (Sproat, 1992) who, like us, has also found that there is a gain in *grouping* rules according to Level.

There is nevertheless broad agreement that, in English, Level 1 and Level 2 are affixational levels dealing with latinate morphology (Class I affixation) and native morphology (Class II affixation), respectively. Level 1 feeds Level 2, therefore native affixation must be attached outside latinate morphology. There is less agreement about the relationship between Class II affixation and native compounding and whether one needs to identify a separate native compounding level. For various reasons we do not have space to go into here, we choose to recognize a distinct native compounding Level 3. Moreover, we importantly recognize a Level 0, which is reserved for non-native (i.e. neoclassical) compounding. In other words, compounding purely involving neoclassical elements must be completed *before* affixation takes place.

Thus, the four distinct levels of our model are:

1. Non-native compounding (neoclassical compounding)
2. Class I affixation
3. Class II affixation
4. Native compounding

Each level has two characteristics: it is cyclic and optional. Cyclicity accounts for recursive structures, i.e. we might find forms such as the following:

[1] Selkirk is cited here only as a reference point: we shall develop our own model as shall be seen.

prefix-II + word + suffix-I + suffix-I[2]
where Level 1 rules apply twice before Level 2 ones.

To apply our model, we used the *Edinburgh Cambridge Morphological Analyser and Dictionary System* (Ritchie, et al. 1992), a component of the *Natural Language Toolkit* developed for the UK Alvey IT Programme. This offers a Koskeniemmi-type analyser (here restricted to handling morphographemic phenomena) and a general purpose unification based analyser which allows the morphologist to express her knowledge via feature bundles of attribute-value pairs in a context-free grammar framework.

Our model is instantiated in our computational wordform grammar as follows. The analysis strategy used by the wordform grammar parser is that of a bottom up chart parser. Each rule in our grammar is marked for level *or levels*. Lexical entries are also marked for level. Thus, a Class I suffix like 'ous' as in 'glorious' is marked for Level 1. Monomorphemic non-affix native lexical entries are also marked by default for Level 1. Thus, if we have the wordform 'glorious', then, in a computational environment, string segmentation, morphographemic rule application and dictionary look-up will yield:

glory((cat noun)(level 1)) and
ous ((cat suffix)(level 1))

These two representations are added to the data structure (a chart). Rules with Level 1 as their domain may now apply, as the basic condition for their activation is present in the chart. They will match with these representations and yield:

glory + ous ((cat adjective)(level 1))

which is still a Level 1 object. This is added to the chart and no further rules apply. This representation may now be generated as a word of English. As Levels are optional, in this case the rules associated with higher Levels do not apply. If we take an underived monomorphemic native wordform, this can be seen conceptually as passing through Levels 1–3, with vacuous rule application. All such wordforms are marked as Level 1 in the dictionary, thus will not be considered, as is correct, by Level 0 rules. The fact that an object is marked for some Level does not block it at that Level: it merely indicates that this is the first Level at which rules may apply – recall that we do not know, in bottom-up analysis, whether e.g. we are dealing with an underived form, until we have finished the analysis, thus we must allow for underived forms to *potentially* combine with affixes or participate in compounding.

Besides the use of four levels in our morphological analysis, we additionally introduced a diacritic feature which explicitly marks *degrees of boundness* for neoclassical roots. Analysis of a corpus of Immunology texts, by various (semi-)automatic methods, produced classifications of neoclassical elements into roots and affixes. Neoclassical roots make up our new category **comb** and display three degrees of boundness: totally free (e.g. *cyst*), partially bound (e.g. *myel-* or *-myel*) and totally bound (e.g. *gen*)[3]. Totally bound forms cannot appear on their own and cannot appear in compound final root position without being suffixed. Partially bound forms cannot appear on their own, but can stand in compound final root position without suffixation. Totally free forms can appear in any position, suffixed or not, and can stand on their own. All neoclassical roots are marked in the dictionary with level information **level 0** and a value for the boundness feature[4]. Those neoclassical elements that we have classed as affixes are dealt with largely at Level 1.

In addition to level ordering and boundness information, other characteristics of our implementation are the use of morphosyntactic head, feature value percolation and *relativized head* (Di Sciullo and Williams, 1987). The important issue for us was to determine whether a wordform is a general language word or a potential term. In our system, we demonstrated how this could be achieved for affixed forms, neoclassical compounds and certain types of native compound. We labelled certain suffixes as typically term forming suffixes on the basis of a sublanguage corpus analysis, attaching the feature value **(wordtype term)** to their dictionary entry (each affix has its own lexical entry). We can then ensure that a suffix with this feature *percolates* its value to the mother node. We used only two wordtype values in our system: **term** and **word**. Besides employing the notions of **head** and **percolation** from Lexicalist Morphology, we also used the notion of **relativized head**. This refinement of the notion of head helped us percolate the relevant information in cases where the morpheme bearing the label **(wordtype term)** was not in *syntactic* head position according to the Right-hand Head Rule.

Our wordform grammar rules generate the following word and term forms involving suffixation (note: prefixation is similar to suffixation thus is not shown): term → word + term_suffix
term → term + term_suffix
word → word + word_suffix
term → term + word_suffix.

Compounding operates in a similar fashion:
term → term + word
term → term + term
term → word + term
word → word + word.

Our use of a unification based word grammar

[2] I, II correspond to Class I affixation and Class II affixation, respectively.

[3] We could have worked with three types of **comb**, however we prefer our current solution as it appears more flexible and expressive to us.

[4] We only use two values for **bound**, however boundness is interpreted by a combination of **bound** and **level** values to give us our 3-way distinction.

then allowed features associated with known terminological elements to be attached to overall wordforms, thus characterising them as potential terms for later assessment by the terminologist. The notion of *terminological head* of a wordform is important in this respect: this refers to the element of a complex wordform which confers term-hood on the whole wordform, which may not be the same as the morphosyntactic head.

As yet, we are only capable of determining terminological status for an unknown word, or wordform containing an unknown morpheme, if it contains a known terminological element (revealed by prior corpus analysis and coded appropriately in the dictionary). For known morphemes there is no problem. By using notions of Level Ordering, we can furthermore impose strong constraints on the form a word (or term) may take. Thus, we can filter and reject as nonwords or nonterms wordforms where an analysis without Level Ordering might postulate a valid wordform of English.

We provide an analysis of a potential term in the following.

Final representation

leukaemia
analysis: 1
(((bound -) (compound -) (level 1) (wordtype term) (category noun))
↑ This is the final representation which postulates that the word 'leukaemia' is a noun, term, Level 1, non compound lexical unit.
L0-to-L1-by-n-or-adj-suffixing
↑ rule name
(((compound -) (wordtype term) (bound -) (level 0) (category comb))
↑ representation of lexical entry *a* data
ENTRY
(leuk ((category comb) (level 0) (bound -) (wordtype term) (compound -))))
↑ lexical entry *a*
(((wordtype term) (bound -) (tie +) (level 1) (makes noun) (suffixes comb) (category suffix))
↑ representation of lexical entry *b* data
ENTRY
(+aemia ((category suffix) (suffixes (noun verb adjective adverb comb)) (tie +) (bound -) (level 1) (wordtype term) (makes noun)))))
↑ lexical entry *b*

Relevant rule

(L0-to-L1-by-n-or-adj-suffixing
((category _noun-or-adj) (wordtype term) (level 1) (compound -) (bound -))
→
((category _n-v-adj-adv-comb)(level 0)(bound -)),
((category suffix)(suffixes _n-v-adj-adv-comb)(makes _noun-or-adj) (level 1)))

We have simplified this example somewhat for exposition: in the EdCam system, a dictionary entry contains fields other than the two shown here (the orthographic form followed by associated morphosyntactic information). Underline denotes a feature variable, whose name indicates the set of possible values taken by the feature. All Level 0 objects have terminological status in our corpus thus we may safely mark wordtype directly on the mother. The feature **suffixes** is used as a subcategorisation frame whose value must unify with that of the affixed object. The feature **makes** indicates what category the affix turns the object it attaches to into. The value of this feature is the one that is percolated, via unification of variable values, to the mother node, to give it its category specification. Subcategorisation and **makes** information is stored in the morphosyntax field of an affix's lexical entry. Our suffixing rules are basically all of this form with variants to take care of suffixation *at* different Levels. There are several rules that take care of mapping *between* Levels 0 and 1 as in the above example. With prefixes, which are typically not category changing, we have a three-way unification. The use of the **compound** feature is used at two levels, the neoclassical level and the native level. Compounds are assigned one syntactic parse only, a left branching one, to avoid problems with overgeneration[5].

A top-level filter takes care of allowing only wordforms that are potential terms to be passed out as results: ((category _any-cat) (bound -) (level _1-2-or-3) (wordtype term)). Note that no Level 0 objects can be so output and that each object must be unbound, have a major lexical category (not **suffix**, **prefix** or **comb**) and be of wordtype **term**.

4 CONCLUDING REMARKS

We have implemented a computational morphological grammar and lexicon that instantiates the abovementioned 4 level ordered morphology of English capable of handling both neoclassical compounding and other complex and simple wordforms in a theoretically satisfactory manner, and furthermore demonstrating that application of theoretically motivated linguistic knowledge enhances term recognition. The identification of this new level is an original contribution to morphological theory and, for the first time, allows neoclassical elements to be integrated in a theoretically satisfactory and elegant way in a model of term and word structure.

Term formation is only one of the factors involved in term recognition. Our research has focussed on

[5] It should be noticed that (compound +) is serving two purposes in this analysis: a) as a strategic value to prevent multiple syntactic analyses of a compound and b) to mark an object as a compounded form. Several features of our grammar are inspired by the simple grammar provided with the EdCam system, however we have substantially altered and added to this featureset and ruleset.

morphosyntactic aspects of term formation insofar as these appear to be more tractable than others, which we have also identified in the course of our research.

Our work has focussed recently on the development of tools for sublanguage linguistic analysis to aid the process of word classification: e.g. to effect inversion of KWIC indexes and to apply techniques of gradual approximation to discover semantic collocations between words (Sekine et.al, 1992a, 1992b).

Future work will further investigate the application of such tools to automatic term recognition and will examine how techniques and research results from the various fields given above in section 3 can be applied to other aspects of term formation and thus term recognition.

5 REFERENCES

Di Sciullo, A.M.,and Williams, E.(1987). *On the Definition of Word.* Linguistic Inquiry Monograph 14, The MIT Press, Cambridge, Ma.

Heid, U., and McNaught, J. (1991). EUROTRA-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications. Final Report. Submitted to DGXIII-B, CEC, Luxembourg.

Mohanan, K.P., (1986). *The Theory of Lexical Phonology.* Reidel, Dordrecht.

Ritchie, G.D., Russell, G.J., Black, A.W. and Pulman, S.G. (1992). *Computational Morphology.* The MIT Press, Cambridge, Ma.

Sekine, S., Carroll, J.J., Ananiadou, S., and Tsujii, J. (1992a) Automatic Learning for Semantic Collocation. *Proceedings of Third Conference on Applied NLP*, Trento, Italy, pp. 104-110.

Sekine S., Ananiadou, S., Carroll, J.J. and Tsujii, J. (1992b). Linguistic Knowledge Generator. *Proceedings of 14th Coling, vol.II*, pp.560-566.

Selkirk, E., (1982). *The Syntax of Words.* Linguistic Inquiry Monograph 7, MIT Press. Ca. Mass.

Sproat, R., (1992). *Morphology and Computation.* The MIT Press, Cambridge, Ma.