[*Machine Translation Today: Translating and the Computer 15*. Papers at a conference…18-19 November 1993 (London: Aslib)]

TERMS ARE NOT ALONE: TERM CHOICE AND CHOICE TERMS

Sophia Ananiadou* and John McNaught‡
* Department of Computing, the Manchester Metropolitan University, John Dalton Building, Chester Street, Manchester, UK, Ml 5GD
‡Centre for Computational Linguistics, UMIST, Sackville Street, Manchester, UK, M60 1QD

This paper assesses the degree to which established practices in terminology can provide the translation industry with the lexical means to support mediation of information between languages, especially where such mediation involves modification. The effects of term variation, collocation and sublanguage phraseology present problems of term choice to the translator. Current term resources cannot help much with these problems, however tools and techniques are discussed which, in the near future, will offer translators the means to make appropriate choices of terminology.

INTRODUCTION

Our objectives in this contribution are:

- To discuss the current relationship between the fields of terminology and translation;
- To assess the degree to which established practices in terminology can be said to provide the translation industry with the lexical means required to support appropriate mediation of information between languages, where mediation nowadays often implies modification;
- To demonstrate how techniques and tools being explored by practitioners in natural language processing could be of help in providing such lexical means, especially in situations where no relevant lexical resources exist.

In what follows, we take a somewhat idealised view of translation — idealised in that although we realise there are many different types of translation and of translators, unfortunately, in this short paper, we cannot hope to take into account every type. We will thus discuss general aspects, trusting that the individual translator will be able to establish relevant links to her working environment. We start by looking at the nature of translation and the relationship of the field of terminology to translation, in turn. Then we consider issues related to variant realisation of concepts in different communicative situations. This leads into a discussion of collocation. The notion of sublanguage is seen to be highly relevant with respect to terminological variation and collocation and we thus next discuss why awareness of sublanguage patterns is necessary in translation. Our interim conclusion is that there are no available terminological resources that offer the translator informative support in relation to term variation and collocation or to sublanguage phraseology.

We then turn to consider which tools and techniques are available commercially or in the process of being developed to aid translators, perhaps indirectly, in coming to grips with variation, collocation and sublanguage phraseology. Here we shall find that there is much cause for optimism.

Translation

We start by making a gross distinction between two kinds of translation: that in which the target text is a *dependent* text with respect to the source text and that in which the target text is a *derived* text with respect to the source text.

Translation is often discussed from the point of view of determining the content, form and purpose of some source language message and then preserving this information in translation. In other words, one is focussing on determining the meaning and form of some source language message, followed by finding a link to an equivalent meaning and form in the target language. It is seen as essential to preserve meaning and form in translation. This is the traditional view of translation and much translation is carried out according to this view, where the resultant target text is linguistically and communicatively equivalent to the source language text. We may say that a *dependent* message is produced under this view: the target depends entirely on the source for its content, purpose, form, etc. Many of the lexical resources available to translators are geared to support this type of translation.

However, when we look more closely at today's translation industry, we see also the increasing need to take into account *modification* of the text in translation, where one may choose — or more commonly be required, via the translation specification — to vary some or all of content, intention and text type, together with other textual and language-dependent parameters. Under this view, translation is seen as mediation, where mediation may, and commonly does, involve modification of the linguistic and communicative parameters given by the source text to produce a related target language text where the relationship is not a direct one. We may say that a *derived* message is produced, when the specification calls for some change with respect to the parameters of the source language text: the nature of the target does not depend directly and entirely on that of the source — a different, but nevertheless related, message is produced. Sager (1) expounds this contemporary view of translation as mediation and modification.

Today, translation is increasingly concerned with choosing how best not only to mediate but also to modify a message with respect to the specification. This implies that translators are constantly faced with a major problem of choice, covering numerous parameters, each of which, in the worst case, may call for a choice to be made among a large set of possible values. We do not claim there is some ideal choice to be made or specified which will result in the perfect translation: there is room for much latitude in any mediation which does not involve strictly laid down forms and modes of expression approaching the formulaic and artificial ends of the language spectrum where there can only ever be one possible translation. In the general case, then, the translator is constrained to choose values from some set of parameters in order to yield an appropriate target text. This set of choices has a direct bearing on how the translator textually realises information at all linguistic levels: pragmatic, semantic, syntactic, morphological, phonological and graphological. Given that the translator must produce, at the end of the day, a string of wordforms in the target language, it is evident that the choice of each wordform (and indeed the dynamic construction of complex target forms by the translator) is potentially conditioned by the set of values that has been identified as the most pertinent with respect to the specification.

This furthermore implies that, in order to be candidates for selection, target wordforms must have particular types of information associated with them. We would naturally expect to judge

whether some target wordform is appropriate by consulting its dictionary entry and determining the (perhaps implicit) set of values associated with it and the closeness of the match between this set and the set of values derived from the specification.

Traditional bilingual dictionaries offer a certain amount of support for dependent translation, as there is an assumption in such dictionaries that text parameters are preserved (or rendered essentially context-neutral), in relation to the main sense equivalents.

However if we are carrying out a derived translation, what aids are there to tell us that e.g. target wordform *X* is the appropriate translation of source wordform *Y* as we are e.g. going from full text to abstract, or report to persuasive marketing document? Bilingual dictionaries do not explicitly mark such information — usually, they simply give lists of quasi-synonymous target words with few distinguishing marks. This is why translators often need to look at monolingual dictionaries in both source and target language where they find much more information on words. However, such information is still largely limited to context-independent senses.

What we have just described is a simplistic model of translation in which choice of wordforms is seen to be conditioned by the values of largely global text or translation parameters, derived from the specification. This furthermore implies a very strong lexical basis to translation and, moreover, essentially an atomic one, in which we combine atoms according to our parameter values. Matters are much more complex, in reality. In particular, local textual or translation conditions may apply; or individual words may themselves condition the selection of other words; or it may be apparently impossible to select one word without selecting another, and vice versa. Thus we are faced with a complex interplay of global and local values. This again is a simplifying assumption, however it will serve for present purposes.

Collocations, phraseological expressions of many kinds, idioms and the like are the typical result of local conditions applying. This, as we know, is a thorny area for the translator, who may manage to construct an apparently reasonable translation, in respecting global parameter values, but then finds the result is not as elegant or idiomatic or natural as might be desired.

We thus find that we need more than information about individual wordforms to help us translate: we need also information about the combinatorial possibilities of wordforms. We need much other information too, but we shall consider only these two types of information here.

The following question then naturally arises: to what extent does the field of terminology give support to translators interested in choosing wordforms according to global or local parameter values and according to combinatorial possibilities? With this question in mind, we now turn to consider terminology.

Terminology

Terminology is often discussed from the point of view of determining the concepts of some special language and establishing equivalence links between the concept systems of corresponding special languages of different natural languages. Furthermore, with its tendency to strive for harmonisation and normalisation in the interests of more efficient communication, terminology is often concerned with establishing names for concepts such that, within some subject domain, in some language, there is ideally but one name for one concept, or alternatively one preferred name. Moreover, terminology places great emphasis on nominal forms.

Given such preoccupations, there is a tension between the concerns of terminology, as classically perceived, and those of translation in an environment where choice of wordforms is an important consideration. It would appear that terminology is concerned with constraining choice, whereas translation expects constantly to make choices. One might argue that terminology helps by offering the translator the right choice, for some defined situation, once the translator has chosen the particular situation: but to what extent is this the case?

At this point, we must be careful to distinguish between theoretical aspects of terminology and practical ones. The practical side of terminology is supported by terminology information systems such as term banks, terminology management packages and the like. These mostly do not reflect contemporary terminology theory. Progressive terminology theory is interested, among other things, in representational issues, such as the design of terminological knowledge representations, and in determining the number, nature and role of terminological relationships used to link concepts in such representations. There are very few such knowledge based systems around. There are considerably more concept-oriented systems in existence, which handle simpler hierarchical structures based on just a few terminological relationships, and large numbers of term-oriented systems which essentially establish relationships between terms as opposed to between concepts. The latter can get by without proper definitions, often preferring contextual examples, whereas definitions play a key role in concept and knowledge oriented systems, as they help fix concepts in conceptual space.

Concept representation issues are highly important for terminology and it appears that knowledge based systems offer greater information possibilities to the user through allowing exploration of conceptual space, than do less sophisticated systems.

Currently, there is great interest in exploring multidimensionality in concept systems (Bowker and Lethbridge (2)). That is, how one may view some concept as belonging to more than one relational system at the same time. Thus, a computer operating system can be viewed as single- or multi-tasking, as portable or hardware-dependent, and so on.

Such work, however, does not call into question the underlying tendency in terminology to prefer one form of a term to all others for some concept. In a sense, it does not need to, as its concerns are of a different nature. However, it is interesting to note that the representational and retrieval mechanisms typically found in modern terminology knowledge bases are eminently suited to handling just the kind of issue that is one of the subjects of this paper: the existence of different forms to represent what is essentially the same concept under different conditions, within the same subject domain. We now look at this topic.

VARIANT FORMS

Although theoretical terminology studies largely ignore contextually-conditioned variant forms of terms, it is not true to say that traditional terminological information systems make no attempt to deal with variants of terms. Let us briefly review the major means employed in terminological resources.

First, we note that the information categories described below do not necessarily all occur in some term resource. There is also variation in the interpretation of each category among term resource. Lastly, we may find that, in some actual resource, categories are non-exclusive: the same or similar information may be given under more than one category.

Context: There are many kinds of context. Our interest here is in the kind of context which attempts to capture a typical or a-typical use of a term, and/or shows (types of) words that typically occur with the term.

Scope note: This narrows down the area in which a concept is typically used, e.g. by indicating it is used only with reference to a particular device and is not a concept found in all devices of that general type.

Usage note: This potentially contains a wealth of information, which supplements that found in the context field. Here we typically find details on level of language (colloquial, formal, etc.), in which circumstances use of the term is mandatory, whether the term is standardised or not, whether it is specific to some company, in which geographical or linguistic region it is used, whether it is a translation equivalent thus deprecated in source monolingual use, and so on.

Synonym: This is usually a reference to another headword considered to be substitutable for the term under study. Such substitution may be qualified according to context of use. This context may, or may not, be clearly set out. If there is no indication of context, the assumption then is that the synonym is substitutable under the same conditions as the entry term.

Source origin and type: These two categories can provide useful information on appropriateness, e.g. by indicating a term as being in use at a certain time period, by demonstrating its use in documents issued by a professional or governmental body, etc., by giving evidence of use in a certain type of text. Source type information is of particular help, all the more so if such information is attached not just to the term itself but also to contexts and definitions.

Variant: This category usually covers the narrow area of orthographic variants, i.e. noting differences in spelling of the entry term.

Abbreviation: Here we may find diverse forms such as abbreviations, acronyms, symbols, formulae and the like which can stand for the entry term. Some terminological resources have separate fields for these forms.

Expanded and reduced forms: These categories typically concern multi-element terms (compounds) and give full(er) forms of the entry term (elements added) or various attested — or indeed potential — shortened forms (elements removed).

It is important to realise that not all resources provide all the above categories. What is more important to realise is that, where such categories are provided in whatever measure, there may be few *dependency* links between the categories. Examples of dependency links are: *synonym* ↔ *usage note; entry term* ↔ *abbreviation* ↔ *source type* ↔ *usage note; source type* ↔ *expanded form,* etc. Let us take the example of *entry term* ↔ *abbreviation* ↔ *source type* ↔ *usage note.* Ideally, we would require source type and usage note information to be attached to both the entry term and to each abbreviation (in its widest sense) recorded: we need to know under which circumstances we can use some abbreviated form to replace the main entry term. Too often, such dependency links are not in evidence, which means that the user is forced into wider searching in the resource in order to discover (or not) the information sought. The lack of such links and indeed of certain categories may be due to several factors: the initial design of the resource; particular original requirements; and so on.

Furthermore, simply because a category exists does not guarantee it will have adequate (or indeed any) information in it. Such information is time consuming to discover and compile. Many term resources contain fullish information for only a limited number of categories, whose recording keeps staff more than fully occupied. Other categories are filled out on an ad hoc basis.

We should also take note of the heterogeneous nature of information in certain categories. The usage note is a case in point. Information is furthermore typically expressed in free text, essentially according to the whim of the terminologist. Ad hoc classifications of e.g. register or linguistic region may be employed. In the worst case, there may be no control over what is expressed in this category nor of how it is expressed. There are some standard classifications that may be adhered to (e.g. ISO country codes) however there is no standard classification for e.g. register or level of language.

The combined effect of heterogeneity, lack of standardised classifications and the use of free text in usage notes is that it becomes impossible to establish precise dependencies, thus burdening the user with laborious searching which may (and often, we suspect, does) prove fruitless.

It could in fact be argued that the terminological resources available to the translator are little better than those conventional bilingual dictionaries which give sets of 'synonyms' for the translation equivalent, with very little, if any, contextual or pragmatic usage indicators. The translator is thrown upon her own knowledge of the target language and its contextual and pragmatic possibilities — even if the target language is the mother tongue, the individual cannot be expected to have complete and instantly recallable knowledge of all contextually and pragmatically determined variations and moreover be consistent in the application of this information.

In the majority of term resources there is then inevitably a lack of detailed and easily accessible information about contextual and pragmatic conditions governing the appearance of terms in texts. This is due partly to design and cost factors, and to the requirements term resources were built to meet, and partly to the hitherto entrenched view among theoreticians (where these had any influence on the design of some term resource) that a concept should have only one preferred realisation no matter the communicative situation. Thus, even if a term resource manager might wish to incorporate pragmatic or contextual information, there has been very little applied research or theoretical work that could aid in the satisfaction of that wish.

Ironically, translators have always been aware of the need for information that would help in decisions about appropriateness. However, they have been ill-served by terminology theory in that respect (although well-served in others), ill-served by published bilingual dictionaries and rather confusedly and frustratingly served by term banks: frustratingly because the information they seek may (or may not) be there, somewhere, but is difficult to track down, and confusedly because there is little apparent concern over the nature and type of information that is recorded in what may well be seen as rather ancillary fields of information.

There are certainly other areas which are much underresearched and which are hardly treated in term resources: a good example is that of expanded or reduced forms. Little work has been done on examining under which circumstances terms appear in expanded or reduced form. There are numerous types of reduction: a form may become reduced over time, mainly for the sake of conciseness and economy of expression; we may find full and various reduced forms for one term together within one text, as ambiguity of reference is excluded given the mutual discourse knowledge that has been built up, etc. It takes careful interpretation to detect reduction, when one is not a specialist, and equally careful interpretation to generate contextually unambiguous (indeed meaningful) reduced forms in the target text. Certain reductions are permissible in certain situations, others are not. One might assume that the head of the expression (in English, typically the rightmost element) is always preserved, no matter what other reductions are performed, however this is not necessarily always the case, as we see in the two following examples, where the heads ('ratio' and 'plate') are omitted in the reduced forms:

'carrier-to-noise ratio'
'C/N ratio'
'C/N'

'maintenance access cover plate'
'maintenance access cover'

COLLOCATION

Up to now, we have discussed mainly problems associated with individual terms: choice of which form is the most appropriate given certain settings for text type, register and so on. We have thus viewed terms as isolated objects whose occurrence is conditioned by various global or local factors.

However, the form a concept takes can be dependent on the co-occurrence of some other term(s) or word: here we enter the realm of what is broadly termed collocation.

Collocation is pervasive in language: letters are delivered, soup is eaten and not drunk, perfume is worn, tea is strong and not powerful, and so on. Linguists have long been interested in collocation (especially British Linguistics — see e.g. Firth (3), Halliday (4), Cruse (5)). There has been much recent work on collocation, especially in computational lexicography and computational linguistics. Investigations of large collections of general language texts have shown how important a knowledge of collocation is for any language user. Church et al. (6) discuss, for example, the similarities and differences in sense between collocations involving 'strong' and those involving 'powerful'. These two words are often defined in terms of each other, yet one cannot simply replace one with the other in combination with some other form, for the most part. In the same spirit, Biber (7) looks at collocations involving 'certain' and 'sure'.

Translators are always searching for "the right way of saying something" — for the right collocation, we might say, in many instances. It is more than a question of being terminologically accurate, it is also a question of formulating a sentence or phrase such that it sounds as if it belongs to the type of language under study. As in general language, so in special languages, in fact even more so: special languages (sublanguages) have remarkably high incidences of collocation, as is apparent from a brief scan of any special language collection. The collocations are here highly distinctive in characterising the language of the field.

Collocation is often seen between verbs and nouns. Here are some examples, drawn from a collection of Immunology texts, where one may easily spot sublanguage collocation at work:

"Phagocytic cells destroy/digest parasitic organisms."
"Cell-mediated immunity defends against viral infections."
"This activity is destroyed by an anaphylatoxin inactivator which digests the arginine."
"An anaphylatoxin inactivator occurs naturally in human serum."
"The disease tends to remit and to respond to chemotherapy."
"The vessels become plugged with thrombi and there is exudation of fluid rich in neutrophils into the surrounding tissues."
"These drugs interfere with the normal metabolic processes."
"Aggressins interfere with normal defence mechanisms"
"An antigen is a molecule that elicits a specific immune response when introduced into the tissues of an animal."
"The normal tuberculin test skin reaction cannot be elicited."
"This hapten binds to the antigen binding site and bonds to ammo acid residues."

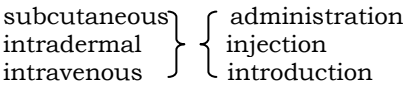
We also note collocation between adverb and verb:

"Compounds of aluminium strongly adsorb protein antigens."

or between adjective and noun:

'rapid death', 'slow death', 'lowered resistance', 'slow infection'.

Note that there can be restricted variation in collocation, as we see in the following:



Collocation in special languages can be markedly different to general language and indeed involve syntactic structures that are apparently deviant with respect to the general language:

"On Monday mornings, cotton and flax workers present with byssinosis."

Here, not only do we have a special collocation involving 'present + disease/ condition/ symptom', but also a special construction 'present + with'. This is one of the most frequent collocations in medically-related texts where the initial state of patients is being discussed and therefore the translator would wish to employ it, rather than generating "the workers appeared with/ manifested the symptoms of/ came looking for treatment for" etc.

As many of the collocates of terms (verbs, adjectives, adverbs) are not themselves considered terms, they will not appear in term resources in any useful sense — they may appear by accident rather than by design, in contexts or other notes, but one would not be able to see an extended set of collocates: think of the case of 'administer/inject/introduce'. Here we further see that we can find a variety of collocates, yet a restricted variety: how can the translator know what this variety is and what forms are acceptable in which circumstances? How can she know that, in a message between consultants, 'present with' is to be preferred, but in a message from consultant to hospital trust board member, the expression 'sought treatment for' might be preferable?

One might say there is no substitute for experience and extensive knowledge of the subject domain, but every translator needs relatively more or less help with choosing or indeed identifying appropriate collocations at times throughout their career.

Frawley (8) neatly sums up the nature of sublanguage, showing the key contribution of collocation:

1. Sublanguage is strongly lexically based;
2. Sublanguage texts focus on content;
3. Lexical selection is syntactified in sublanguages; thus
4. Collocation plays a major role in sublanguages;

5. Sublanguages demonstrate elaborate lexical cohesion.

The particular structures found in sublanguage texts reflect very closely the structuring of a sublanguage's associated conceptual domain. It is the particular syntactified combinations of words that reveal this structure. Techniques which allow us to establish measures of association between the wordforms of sublanguage texts can reveal much about collocational behaviour and semantic classes.

The lesson to be drawn from study of sublanguage texts is not only that collocation is important, but that it is essential as a communicative device: it carries greater communicative weight than in general language.

Up to now, we have avoided defining what we (or others) mean by 'collocation'. Unfortunately, definitions of collocation are numerous and varied. Some researchers include multi-element compounds as examples of collocations; some admit only collocations consisting of pairs of words, while others admit collocations consisting of up to, say, five or six words (there may be intervening material); some emphasise syntagmatic aspects, others semantic aspects. The common points regarding collocations appear to be, as Smadja (9) suggests: they are arbitrary, they are domain-dependent, they are recurrent and lastly the occurrence of one word (or more) strongly influences the occurrence of others.

It is not a goal of this paper to offer yet another definition of collocation. However, what we can observe is that there is, from a terminologist's point of view, little advantage to be gained in viewing multi-element compound terms as collocations: they are terms, with all that this implies. The fact that their elements may occur in combination may be useful as one of the guides to recognition of unknown compounds, however to characterise multi-element compound terms as collocations is, in our opinion, to fail to recognise their special nature as terms. Multi-element compounds may however be quite well characterised as collocations in general language: but that is the subject of a different paper.

In our view, there is equally little to be gained from applying straightforward frequency based techniques which will deliver, as collocations, among other things, somewhat trivial combinations of words, or combinations of trivial words. We might find combinations which have low frequency of occurrence yet are still highly significant as collocations.

Entire phrases, sentences and paragraphs have also been treated by some as collocations. This links in with our comments on collocation in sublanguage where we see highly syntactified structures are predominant. As soon as we expand our view of collocation from, say, pairs of words to numerous words combining and cooccurring, it is then a small step, in sublanguage texts, to consider entire text units as collocations. However, we should be careful to distinguish between:

1. Fixed phrases such as idioms;
2. Formulaic repetition in certain text types dealing with certain subjects, where certain pieces of information are relatively constant;
3. Instances of sublanguage specific sentential or phrasal templates.

We have nothing to say, here, regarding idioms. Examples of formulaic repetition include regularly

issued bulletins where, for example, titles, captions and whole pieces of other text stay constant for every issue at some time period, although they appear to change from day to day or hour to hour:

"Weather report for the 12 hours ended 6pm Tuesday"
"Weather report for the 12 hours ended 6am Wednesday".

Such repetition is also well known in e.g. administrative and legal texts. There is however a matter of degree of mandatoriness to consider: a formulaic statement that is changed (for example in translation) in a weather bulletin may cause less problem than one changed in a legal document. This indicates that text type, subject matter, content and intention must be taken into account when one wishes to say anything meaningful about formulaic repetition.

As for instances of sublanguage specific sentential or phrasal templates, here we may refer once again to the syntactified nature of lexical selection in sublanguages. Certain combinations crop up again and again. Above, we considered all of these as collocations. However, if we continue to do so, we will miss valuable generalisations: we will miss the patterns inherent in the text. Work in this area has been going on for many years, largely ignored by the natural language processing community. For example, medical sublanguage has been the subject of a long-term effort at New York University (Sager (10)). Numerous patterns and templates have been identified and used to build information systems and other natural language applications. These patterns are typically centred on verbs and have thus much in common with attempts by general linguists to describe the argument structure or frames of verbs. However, the potential number and nature of the verbal arguments in sublanguage frames are often quite different to those discussed by general language linguists. It is often argued by detractors that medical sublanguage lends itself remarkably well to such interpretations. However, we find the same kind of patterning in other sublanguages. For example, a satellite telecommunications verb frame might be:

TRANSMIT[SIGNAL_SOURCE,SIGNAL,FREQUENCY,SIGNAL_DESTINATION,MEDIUM]

"The satellite transmitted a test signal on 100MHz to the ground station through free space."

A verb frame for the same sentence constructed by a general linguist might be:

TRANSMIT[AGENT,PATIENT]
(where patient here is to be read as 'entity undergoing some action')

Thus, for a general language linguist, all the prepositional phrases would constitute adjuncts (or circumstancials) which would be seen as having little central role to play, whereas for the sublanguage specialist, the prepositional phrases are critical: they are arguments (perhaps optional, perhaps not) of the sublanguage verb and serve to indicate links between concepts governed by the verb transmit. Note that the following would never be construed as a sublanguage usage of transmit, even though it has apparently the same structure, the same functions and dependencies:

"The European Court transmitted its brief opinion to the British Government on a low-loader lorry through Belgium".

One might reasonably detect a collocation here ('transmit an opinion') however the difference between the two sentences lies exactly in the terminological density and patterning of the sublanguage sentence, where domain concepts are linked together by the sublanguage verb to form a meaningful conceptual statement: meaningful in the sublanguage. Note also that variation in sublanguage

sentences is usually quite restricted:

? 'It was through free space that...'

? 'It was on 100MHz that...'

? 'To the ground station, through free space, was transmitted...'

Also, sublanguage nonsense can be obtained which might appear quite acceptable in general language:

sublanguage sense: "We washed the polypeptides in hydrochloric acid."

sublanguage nonsense, general language sense: "We washed the hydrochloric acid in polypeptides."

This example is due to Harris (11), who points out that we cannot exclude the second sentence from the general language, where some metaphoric meaning could be intended. However, this sentence would never occur in sublanguage texts.

Such restrictions, not just syntactic but also lexical, morphological, etc., mean that there are fewer possibilities for expression in sublanguages. Combined with great conceptual, terminological density, this further means that sublanguage texts tend to use the same means to talk about the same things a lot of the time: to the general linguist, this then gives the impression of collocation at work. However, we may go a step further and say that there are underlying patterns and templates at work that characterise the semantic, conceptual nature of sublanguages. Such a realisation then allows us to collapse whole series of apparently different patterns into similar patterns: at the collocational level, such generalisation would be missed. For example, in the NYU research referred to above, words (largely terms) were grouped successfully into word classes, which then give rise to semantic classes that can be used to build frames. A simplified example is the 'General Medical Management' frame:

[INSTITUTION PATIENT MANAGE.VERB]

where

INSTITUTION has as members: 'cardiology', 'clinic', 'casualty', 'hospital', 'lab', 'outpatients',...

PATIENT has as members: 'patient', 'pt' (abbreviation), 'she', 'he', ...

MANAGE_VERB has as members: 'admit', 'diagnose', 'discharge', 'evaluate', 'transfer',...

This then allows one to recognise or synthesise sentences such as:

"patient was admitted to hospital"
"pt was transferred to outpatients"

and so on: the kind of phraseology that occurs time and again in medical reports and that should thus be desirably reflected in a translation.

In the foregoing sections, we have examined issues of choice concerned with terms: we have explicitly or implicitly considered choice at the authoring stage as well as at the translation and generation stages. We have seen that choice, in sublanguage texts, involves being aware of communicative context, of text type, intention, translation specification and so on. A concept may be realised in different forms, related or not to some base form, depending on such factors of the translation environment. We have also seen that the choice of certain terms and their syntagmatic positioning can be and often is highly dependent on the occurrence of other terms. This latter phenomenon

can be approached as involving collocation; it is a special kind of conceptual collocation, though, as there are clear underlying patterns detectable, which allow us to describe, in abstract form, many apparently different collocations as manifestations of a few (often simple) constructions. It is this abstract view that, in turn, allows us to choose, within acceptable parameters, appropriate surface realisations. In other words, we can know the preferred modes of expression and be able to introduce variety in our target text: variety which remains within the bounds of acceptability with respect to the particular sublanguage we are using.

It is all very well to discuss notions of choice — this is a topic about which translators need to be told few facts. We may have helped shed some theoretical and practical linguistic light on certain issues, to do with terminology and sublanguage. However, what is of key interest is: how can one discover and exploit such information in practical ways? What tools and resources are there to help the translator make the appropriate choice in some circumstance? We turn now to consider these points.

TOOLS AND TECHNIQUES

First of all, we may note once more that there is a great lack of information of the kind we have been discussing — there are vanishingly few lexical resources that store such information in a formal, easily searchable and retrievable way. We have seen that most term banks cannot help out. Contextual variation in terminology is not handled well in term banks. As for collocation, we have in fact come across only one term bank which has been explicitly constructed to store collocational information in a formal way, for multilingual purposes. This is a term bank at Krupp Industrietechnik GmbH, based in part on Hausmann's theory of collocation (see Freibott and Heid (1990) for a description of this bank and e.g. Hausmann (13)).

If the translator cannot find the information in term banks or dictionaries then she must look to means to enable her to discover such information, to tools or techniques that can be directly or indirectly used. In the last few years, large scale processing of texts, in the form of ad hoc collections or deliberately designed corpora, has become widespread in computational lexicography and natural language processing. The reasons for this need not detain us. User requirements of corpora for natural language processing are discussed in McNaught (14). Briefly, in order to build better natural language processing systems or dictionaries, we need to process large bodies of text to discover facts about language. Much of this work is done by applying various tools, mainly relying on statistical and probability-based techniques.

We will now look at several types of tool and techniques. Of necessity, our discussion will be brief: our aim is not to offer an exhaustive catalogue of potentially useful tools, but to draw the attention of translators to types of tool that offer help with problems of term choice, collocation and sublanguage phraseology.

'Key Word In Context' Tools

A clearly useful type of tool, of which there are many instances on the market, is that which produces a Key Word In Context (KWIC) output. Such tools have been around for many years and form a standard utility for anyone interested in processing text to discover, in a limited way, lexical and collocational regularities and associations. In passing, we mention that inverse KWIC tools also exist: by a simple transformation, these show different words that appear in the same context.

Translation Memory Systems

We will now look in greater detail at a more sophisticated type of tool that has been available on the market for some time now and consider the help that translation memory systems can offer.

Translation memory systems have become popular recently, with several systems on the market. These rely on the existence of previously translated texts i.e. on corresponding source and target texts. They firstly process these data in order to, for example, align phrases and establish links between corresponding words and phrases (one may also, typically, build up a translation memory incrementally as one goes about translating). They exploit pattern matching techniques to discover phrases in their structured memory which are identical or close to some phrase the translator has selected in the source text she is working on, and then display the associated translations. The translator can then choose to incorporate what is offered, or not. Maximum benefit is gained when texts being translated are highly similar to previously seen texts. This is the case for successive versions of a manual for some device, for example, when the bulk of the material does not change from version to version. Systems often offer, in addition, integrated terminology management packages. Systems on the market include TM/2 (IBM), Eurolang Optimizer (SITE/Eurolang) and Translator's Workbench (Trados) with their attendant utilities. To what extent can translation memory techniques help with variant term choice or collocation and phraseology choice? Insofar as one is able to look at patterns in one language and translationally (partially) equivalent patterns in another, they do indeed help. If one is able to specify detailed control information for archive texts (type of text, author, date, company, subject domain), then such information can be used to impose a ranking on retrieved matches. If the system can exploit an associated terminology resource, then possibly the translator can browse through variant term forms for both source and target language segments. One cannot, however, as yet expect too much of such systems. It appears that integrated terminology management packages are used more often than not by translators themselves to record term-term correspondences that the translation memory proper has not yielded.

Furthermore, one must be careful in distinguishing between statistically or probability based pattern matching and linguistic or interpretative pattern matching. A translation memory has no knowledge that a form may be a term, unless it is told so (e.g. via explicit annotation in an associated terminology resource — these might more properly be called wordform resources). All it sees is patterns standing in some alignment relation; it can determine closeness of match to some given string according to probabilistic, statistical and positional information. The selection of the string to match is in the end up to the user: a string can be any arbitrary sequence of characters, in effect. Also, the system has no real knowledge about the nature of the relationship between source and target segments, beyond the fact that one has been used as a translation of the other. This is not to say that the target segment is an appropriate translation in any way. If, for the sake of discussion, previously translated texts were translated by a person who had little knowledge of the terminology and phraseology of some sublanguage, then it will be segments of the result of such translation that the user will see. In such a situation, if the user has likewise little knowledge of that sublanguage's terminological and phraseological behaviour, then she will get no true help from the system and, if she accepts what is proposed, will merely propagate an inappropriate translation. Thus, the usefulness of translation memories is directly linked to the quality of the previous translations they are dependent on. We certainly do not deny their clear utility, we merely point out that, if one is using one's own previous translations as a source of information, this will be helpful only if one is happy with the quality of one's previous work and has some means of ascertaining its appropriateness to the task in hand.

This is not so much a criticism of translation memories, which can indeed help greatly in the translation task. It is more a reminder that the functionality of any tool must be carefully studied with reference to the translation environment it is being considered for (and also that existing environments could well bear re-appraisal in the light of tool functionalities on offer). For those who are interested in adequacy evaluation of translation memories and indeed translators' aids in general, we recommend study of EAGLES (15). The CEC sponsored Expert Advisory Group for Language Engineering Standards is working on promotion of de facto standards in a number of areas, including adequacy evaluation of translators' aids. At the time of writing, a substantial draft report is available, for comment by and feedback from the community. It is intended to publish recommendations for de facto standards in the Autumn of 1995. For further information, the reader is advised to contact the EAGLES Secretariat, Consorzio Pisa Ricerche, Piazza A. D'Ancona 1, 56127 Pisa, Italy.

Prototype Tools and Techniques

In the following sections we shall discuss tools that have not as yet appeared on the market: the techniques on which they are based are either of recent date or have been recently adopted and adapted from other areas, chief among which is the area of information retrieval. In many cases, we are thus dealing with prototype systems or with techniques which could be usefully applied to our problem-area after further development. However, other techniques could find rapid application with a minimum of work by a competent programmer. Thus, while we realise the translator or terminologist may not be able to go out and pull such tools off the shelf, we nevertheless discuss them here as:

- Translation and terminology organisations may be interested in commissioning implementations of the more straightforward techniques;

- Several projects throughout the Union and in the USA have been launched to provide various kinds of corpus exploration tools.

Regarding the latter, these are as yet at early stages of development but we should see the results of these projects being eventually commercialised. Thus, it is good to know in advance of their existence. The CEC in particular has been instrumental in supporting research into the development of corpus processing tools, in the framework of the Linguistic Research and Engineering programme, run out of DG XIII in Luxembourg. Cencioni and Klein (16) give synopses of current LRE projects, which are conducted on a collaborative basis between industry and academia, sponsored by the CEC. The most relevant of these projects in the light of our topic are: DELIS, COMPASS, MULTEXT, TRANSLEARN, TRANSTERM and GIST. There are numerous other LRE projects dealing with other aspects of language engineering that may equally interest the reader. In the UK, the Speech and Language Technology (SALT) programme of the EPSRC/DTI has supported collaborative industry-academia projects such as the British National Corpus Initiative, ACRONYM (collocation retrieval of thesaurally related items) and DRAFTER (assistant for technical writers to produce drafts in English or in French). The reader is advised to contact Dr Peter Lee, Department of Trade and Industry, 151 Buckingham Palace Road, London, UK, SW1W 9SS for further details of these projects.

In the USA, there is a major corpus project at the University of Pennsylvania, which as well as building corpora is developing numerous tools to explore them (Marcus et al. (17)). Almost every corpus project is engaged in building tools to process their texts, there being few suitable tools on the market.

The spin-off from all these projects should therefore be significant in terms firstly of tools to explore corpora or text collections and, eventually, grammars, dictionaries, resources and full-blown natural language processing systems and other aids built on the results of all this corpus work.

Our discussion will remain general as we wish rather to point out potentially useful techniques. Fortunately, the contribution by Erlandsen in this volume describes one of the few commercially available tools able to offer flexible means of processing data to yield substantial information on cooccurrence phenomena. Hopefully, our comments will enable the reader to appreciate the general nature of the type of technology involved in such tools.

Tools and Techniques for Collocation

As we have intimated, much research has been going on in this area recently, inspired mainly by statistical and probability based algorithms found in information retrieval.

Favourite techniques involve the use of measures of similarity such as *Mutual Information,* or of dissimilarity such as *t-score*. Church et al. (6) provide a clear and informative discussion of the use of these two measures in lexicography. Further exposition is provided in Charniak (18). Mutual Information operates with pairs of words as follows: it considers how probably one might come across the pair together, how probably one might find each member of the pair on its own (without the other, i.e. by chance), then compares these probabilities and yields a value which denotes the strength of the association. One may successfully determine strong associations, uninteresting associations and pairs whose members are essentially in complementary distribution. One may thus rank all combinations of some word with all others, determine a threshold value and consider associations above that threshold to be relatively strong for that word.

With measures of similarity, it is not so easy to determine the difference between two words which are close in meaning, by looking at the pairs which each participate in. That is, it is easier to find evidence to support some hypothesis than to find evidence against it: it is difficult to determine what words do not occur after some given word. We suffer from lack of evidence or uncertainty about whether our evidence is adequate. Our lack of evidence might simply be due to not having processed enough data or to having used the wrong technique. Thus, Mutual Information has its limits.

However, we can employ a measure of dissimilarity, such as t-score, to help us determine to what degree closely related words differ. This measure utilises the notion of the *null hypothesis* (i.e. that there is no difference): we first compare the probability of word *X* occurring with word *Z* against that of word *Y* occurring with word *Z.* Then we ask what likelihood there would have been of observing any difference between the probabilities if the difference had in fact been zero. If we find this likelihood to be significantly low (less than 1 chance in 20) then we can reject the null hypothesis.

Mutual information and t-score give different, but complementary, results. They can only be used to examine the association between pairs of words, however they can nevertheless give very useful information about collocation of not only nouns and nouns, or nouns and adjectives, but also, for example, verbs and prepositions.

Programs to apply these measures are straightforward to write, especially in environments which offer powerful utilities as standard (as does, for example, the Unix™ operating system). Church (19) is a brief tutorial containing short yet complete and fully operational programs (rarely over 1 page) to implement these measures and other similar ones, which was given to a largely non-computational audience in the interests of encouraging wider use of these techniques.

Even more interesting results may be obtained if the objects being compared are words labelled with part of speech — automatic morphosyntactic tagging of large text collections is an entire activity in itself that we gloss over here: most of the corpus-related projects we have mentioned are developing or have developed such tools. Previously, we were considering raw wordforms and thus could not distinguish between, 'bank'$_{NOUN}$ and 'bank'$_{VERB}$, or 'to'$_{PREPOSITION}$ and 'to'$_{INFINITIVALMARKER}$ for example. However, once we know the part of speech of wordforms, we can then produce more precise information about the specific behaviour of wordforms and be able to distinguish between homonyms.

It is also possible to look for collocations on the basis of syntactic structure: there are tools offering skeletal parses (syntactic analyses) of texts which trade accuracy for robustness and rapidity — as we are interested mainly in gro ss syntactic structure, then their output is valuable. We can thus determine, by applying statistical techniques on the results of such parsers, the typical objects of certain verbs, or the typical verbs of certain subjects, and so on.

Church et al. (6) and Smadja (9), among others, demonstrate how statistical techniques may be combined with linguistic information to yield collocational information. The methods each use are different. As we saw, Church and his colleagues work with combinations of two words; Smadja's work, in addition, offers the possibility to look for collocational behaviour in combinations involving up to thirty words.

It should be noted that the two measures we mentioned are not panaceas, either singly or together. Many statistical techniques of this type are affected by the sparse data problem, for example, or yield certain amounts of rather odd results. We can attempt to mitigate these effects by introducing linguistic filtering, e.g. via tagging text with part of speech labels, however such effects will always remain to some extent, depending largely on the nature of our texts. It should furthermore be noted that most of the work in this area has been concerned with processing large scale text collections of general language. This is not to deny the usefulness of the techniques discussed for terminology: we simply warn the reader to be sensitive to the current general language orientation of the technology and yet not be dismissive of it because of that orientation.

Term Recognition Tools and Techniques

In our discussions so far, we have quietly glossed over a very important aspect of term choice: how to know in the first case that we are dealing with terms as opposed to general language words. In order to know that some form is a variant of a term, if we have no prior record of it, we must be able to recognise it as having terminological status. In order to detect special language collocational behaviour involving nominal terms and verbs, we need to also know that the nouns we are investigating are terms, especially if we have no prior information on these forms. It might be supposed that the statistical and probabilistic techniques we have looked at could help in the recognition of terms. To a certain extent they do, however often forms that are terms are not picked up by associative techniques, as no evidence can be found to propose a strong association among the elements of e.g. multiword compound terms. Straightforward counting of frequency of occurrence can help (on the hypothesis that frequently occurring forms should represent the most important concepts of specialised texts) but is also misleading: one finds elements being proposed as terms that clearly do not have such status.

Recently, there has been an increase in research into this entire area. Daille et al. (20) propose an approach combining statistical and linguistic techniques, applied to aligned texts (original plus translation), to discover compound terms, where the statistical techniques used are sensitive to both frequency and association characteristics of the data. A disadvantage of this work, as with all such

work involving aligned texts, is that the quality of the translation is always in doubt, thus results must be interpreted with caution. A good overview of the problems of extracting multiword compound terms is given by Lauriston (21).

The particular form that terms take, in various text types and subject areas, is furthermore critical to their recognition as terms. Each domain has its preferred methods of term formation. It is important to have knowledge about term formation possibilities and to know how formations may be affected by a change of register, of text type, of communicative situation and so on. Among the various types of term formation are: derivation, compounding, back formation, borrowing, simile, conversion, compression, and so on. Ananiadou (22) investigates how linguistic knowledge of term formation can be used to drive a term recogniser. Such knowledge is also highly useful in aiding the translator or terminologist in the synthesis of terms and in helping her to decide how to realise a concept in some context. What is clear from research in this area is that certain types of term formation are quite intractable at present, from the point of view of trying to recognise them in running text. Also, it is clear that even successful processing can only hope to propose potential occurrences of terms. Human interpretation must in the final analysis be brought to bear to decide whether a form is indeed functioning as a term. The aim of automatic term recognition is then to attempt to recognise all potential, rather than actual, terms in a text, hopefully including all actual terms within the set of potential terms discovered, while excluding forms deemed not to be terms.

Text Type Analysis

Regarding computational analysis of text type and communicative contexts, this too is an area that is attracting greater interest from researchers. As yet, much of the work is focussed on general language. Important results in this area are due to Biber (7), who moreover discusses how his techniques could be extended to specialised texts. Translators and terminologists can look forward to further developments in this area which will directly affect their work, as these will provide the means to discover information regarding the functional nature of text types, the communicative role of various modes of expression and so on.

CONCLUSION

The reader can thus appreciate that there is much research going on into applying various techniques to the processing of collections of texts to yield information about the behaviour of wordforms. Experience gained from using the prototype tools we have described will undoubtedly feed into the construction of commercially available tools to aid in the extraction of knowledge about how terms behave in context. This can only be to the benefit of translators and terminologists.

Eventually, term resources will hopefully offer the means to store, and search for, variant termforms, collocations and sublanguage phraseology. However, at present one can only indulge in self-help, although the tools and techniques described are a definite aid and their relevance to the translator should not be ignored.

In closing, we wish to make a final methodological point. There is certainly a strong temptation to process paired source language original and target language translated text, given that quantities of such 'parallel corpora' exist. There is apparently an equally strong belief that such processing will yield good quality terminological data, collocational and phraseological information that will then be of use to translators. We are not so convinced of this, as we have hinted at earlier, as one cannot be at all sure of the quality of the translation and particularly whether it did indeed respect the target language constraints on phraseology, collocation and term choice. Furthermore, the processing

of paired texts will not help overmuch with translation situations where modification of the message is called for: as far as can be seen, most parallel corpora are of the dependent translation type we mentioned at the beginning of this paper.

We believe that, if one wishes to arrive at the best possible information on collocational, phraseological and terminological behaviour, it is paramount to process original texts in the target language, rather than translated texts. The tools and techniques we have discussed are entirely usable to this end. Arntz (23) reminds us that, in comparative terminology, one does not work with translations, but with original texts in each of the languages under study, in order to determine firstly the conceptual and terminological system of each language independently and only subsequently to establish mappings between these. This is a methodology that should be adopted at all levels of terminological investigation, thus applicable to term variants, collocations and phraseological behaviour. It is not an easy task to work in this manner but the results are bound to be of higher quality than if we had worked with translated texts. After all, the translator wishes to determine how information should be expressed, given some communicative situation, in the target language. Such knowledge is really only to be found in original texts of the target language and original texts of the source language.

REFERENCES

1. Sager, J.C., 1994, "Language Engineering and Translation", John Benjamins, Amsterdam.

2. Bowker, L., and Lethbridge, T., 1994, Terminology and facetted classification: applications using CODE4. In Albrechtsen, H., and Oernager, S., eds., 1994, "Knowledge Organization and Quality Management", Indeks Verlag, Frankfurt, 200-207.

3. Firth, J.R., 1957, A synopsis of linguistic theory 1930-1955. In Palmer, F., ed., 1968, "Selected Papers of J.R. Firth", Longman, Harlow.

4. Halliday, M.A.K., 1966, Lexis as a linguistic level. In Bazell, C., Catford, J., Halliday, M., and Robins, E., eds., 1966, "In Memory of J.R. Firth", Longman, London.

5. Cruse, D.A., 1986, "Lexical Semantics", Cambridge University Press, Cambridge.

6. Church, K., Gale, W., Hanks, P., and Hindle, D., 1991, Using statistics in lexical analysis. In Zernik, U., ed., 1991, 'Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon", Lawrence Erlbaum Associates, New York, 115-164.

7. Biber, D., 1993, Using register diversified corpora for general language studies, Computational Linguistics 19:2, 219-242.

8. Frawley, W., 1988, Relational models and metascience. In Evens, M., ed., 1988, "Relational Models of the Lexicon, Cambridge University Press, 335-372.

9. Smadja, F.A., 1993, Retrieving collocations from text: Xtract, Computational Linguistics 19:1, 144-177.

10. Sager, N., 1986, Sublanguage: Linguistic phenomenon, computational tool. In Grishman, R., and Kittredge, R., 1986, "Analyzing Language in Restricted Domains", Lawrence Erlbaum, Hillsdale NJ, 1-18.

11. Harris, Z., 1982, Discourse and sublanguage. In Kittredge, R., and Lehrberger, J., eds., 1982, "Sublanguage: Studies of Language in Restricted Semantic Domains", Walter de Gruyter, Berlin, 231-236.

12. Freibott, G., and Heid, U., 1990, Terminological and lexical knowledge from computer-aided translation and technical writing. In Czap, H., and Nedobity, W., eds., 1990, "TKE'90: Terminology and Knowledge Engineering", Volume 1, Indeks Verlag, Frankfurt, 522-535.

13. Hausmann, F.J., 1985, Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In Bergenholtz, H., and Mugdan, J., eds., 1985, "Lexikographie und Grammatik", Lexicographica, Series Maior 3, 118-129.

14. McNaught, J., 1993, User needs for textual corpora in natural language processing, Literary and Linguistic Computing 8:4, 227-234.

15. EAGLES, 1994, "Evaluation of Natural Language Processing Systems", Draft Report, EAGLES Document EAG-EWG-PR.2, Consorzio Pisa Ricerche, Pisa.

16. Cencioni, R., and Klein, E., eds., 1994, "Linguistic Research & Engineering: An Overview", DG XIII E-4, CEC, Luxembourg.

17. Marcus, M., Santorini, B., and Marcinkiewicz, M.A., 1993 Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics 19:2, 313-330.

18. Charniak, E., 1993, "Statistical Language Learning", The MIT Press, Cambridge, MA.

19. Church, K.W., 1994, Statistical Tools — Unix™ for Poets. Proc. Language Engineering Convention. Paris, July 1994, EC2, Paris.

20. Daille, B., Gauissier, E., and Lange, J-M., 1994, Towards automatic extraction of monolingual and bilingual terminology, Proc. COLING-94. 515-521.

21. Lauriston, A., 1994, Automatic recognition of complex terms, Terminology 1:1, 147-170.

22. Ananiadou, S., 1994, A methodology for automatic term recognition, Proc. COLING-94. 1034-1038.

23. Arntz, R., 1993, Terminological equivalence and translation. In Sonneveld, H., and Loening, K., eds., 1993, "Terminology: Applications in Interdisciplinary Communication", John Benjamins, Amsterdam, 5-19.