[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

The Use of Sublanguages in Machine Translation

Effie ANANIADOU*

July, 1990

1 Introduction

In this paper we will examine the impact of the notion of sublanguage for MT. We will see also to what extent sublanguage has had an impact on linguistics and NLP. We will define the notion by identifying the relevant properties useful for MT. In this account we will be general enough to talk about MT and not about Eurotra specifically. The reason for this is to show why we thought in the first place that sublanguage techniques could be of use to our project. The remaining sections will provide support to our initial hypotheses and claims. Then we will assess how sublanguage can be applied to MT.

First we will examine to what extent linguistics, NLP and MT have made any use of the notion of sublanguage.

Linguistics typically concentrates on 'general language', on competence as opposed to performance. NLP has different goals from linguistics, is concerned at any one time with a limited part of the language, the actual use of language. Typically, an NLP project deals with a limited subset of a natural language. Linguistics bases its descriptions on portions only of the natural language while we could argue that NLP's goal is to provide significant generalisations. What is significant about NLP is that it uses linguistic information, or other information from the field of e.g. AI (in text understanding and in making of inferences), in a practical and convenient way.

*Centre for Computational Linguistics, UMIST, Manchester M60 1QD, UK. E-mail: effie@uk.ac.umist.ccl

2 Towards a definition of sublanguage

Language is viewed as a variety of interacting systems having a general and special purpose. Sublanguages are associated with special purpose, reflected in the use of the term 'special languages' in preference to 'sublanguages' in Europe. Sublanguages are defined as:

> *... intersecting subsystems which overlap general language and are therefore variously dependent on it.*
>
> (Sager et al., 1980:65)

The relationship between sublanguages is not one of total disjunction - rather they form mutually exclusive, though overlapping sets based on divisions of knowledge of a community. The notions of intersection, overlapping, subsystem are important in defining sublanguages and we will see that they are important in practical terms for MT. We will examine first the relation between general language (GL) and sublanguage (SL).

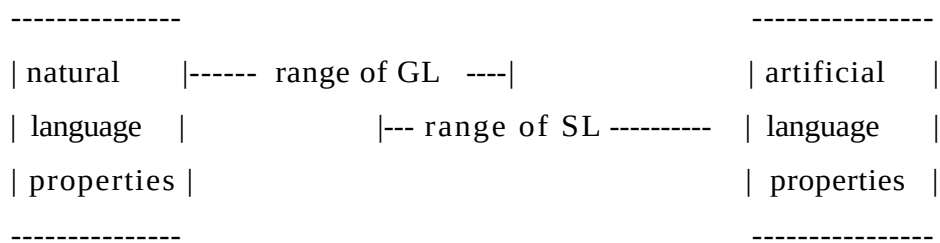
(i) there are degrees of intersection between a SL and GL

This statement emphasizes the fact that the boundaries between varieties of language are not clear cut but fuzzy. There is a similar blurring of boundaries between varieties of natural language and artificial (formal) languages, and sublanguages may extend either side of this boundary, or be indeed wholly on one side or the other:

> *If we were to apply the same criteria for complexity to ... formal languages and to various 'natural' sublanguages of English, members of the sublanguage list would probably not all lie beyond the point on the scale corresponding to the most complex of the formal languages.*
>
> (Lehrberger, 1986:36)

The same idea was mentioned by Sager et al. who view sublanguages not as constituting a class of their own at one end of a dichotomy, but as being distributed over a continuum, whose end points are GL on the one hand and strictly artificial language on the other. The borderline between natural and artificial language lies at the point where natural language loses the ability to be its own metalanguage. Harris (1976:276) stated that 'subject-matter sublanguages' are distinguished partly by the fact that they do not contain the sentences of their own metalanguage. However, the natural language contains its own metalanguage, namely the grammar which describes it.

The relationship between GL and SL, and between natural and artificial language, may be characterized by the following diagram:



(from Sager et al., 1980:41)

(ii) the grammar of a SL is not necessarily included in the grammar of the whole language

This statement is an important one for sublanguage studies, as it endows sublanguage with its own status. Complexity however is introduced as the 'normal' case is for the grammars of GL and a SL to intersect. Thus a SL grammar is not just a subgrammar of the standard language. Some rules of a SL are not satisfied by sentences that are part of the GL. Also, conversely,

(iii) there are rules of the GL which do not apply in a particular SL

A typical case, to exemplify, would be those rules regarding the use of colloquial forms, idioms etc., or other phenomena which have very restricted realisation.

(iv)

> *... the SL deals with an organised, if not closed part of the real world, whereas the whole language imposes only the broadest structuring upon our perceptions of the world.*

(Kittredge, 1982:235)

This statement contains much of interest for NLP. It has often been said (and demonstrated) that NLP systems perform better when various restrictions are placed on e.g. the data, the grammar coverage, etc. In many cases, such restrictions have been carried out in a relatively ad hoc or non-portable fashion. The above statement emphasises that we can hope to evolve better motivated restrictions by discovering and characterising the organised nature of some sublanguage (manifested e.g. through its text types). The restrictions that are useful to us for NLP are indeed inherent in some sublanguage, as sublanguages are partially characterised by their closed nature.

Closure is one of the main properties of SL. There are degrees of closure of SL but in contrast to the GL sublanguages are more or less closed. Those sublanguages demonstrating a high degree of closure are clear candidates for successful NLP.

(v) The property of closure is associated with the poor creativity of expressive means compared to GL

It should be noted that a sublanguage may show highly productive processes at work, especially in term formation. This is not to be confused with creativity of expressive means. GL utilises a wide variety of expressive means. SL on the other hand tends to select a relatively small number of means. Therefore we are here concerned with the range of means. If we can ignore several or many expressive means, as they are simply not utilised in some sublanguage, then we can omit rules for dealing with their associated phenomena, hence cutting down on the complexity of our descriptions and on the possibilities for overgeneration. It may be the case that a sublanguage shows high productivity for some term formation method or for some syntactic construction. This then shows that the SL has poor creativity of expressive means in comparison to the GL.

(vi) A SL is relatively incomplete as it describes only a limited amount of reality

Incompleteness here is to be understood in relation to GL. It is impossible for a SL to describe as great a part of reality as GL. If it did so, it would no longer be a SL, but GL itself. SL specifically deals with a closed universe, therefore is inherently incomplete with respect to GL. This is again an interesting feature for NLP as it offers the possibility of delimiting the universe referred to, hence allowing hopefully more successful attempts to be made at characterising e.g. semantic relations, real-world knowledge, valid inferences, etc.

2.1 Factors that give rise to SL

We mentioned above that SLs have special purposes. By this broad characterization we mean the whole set of situations, contexts that are part of the SL instantiation. Defining the 'purpose' of a text brings us close to discourse considerations. Any model of SL should take into account discourse. Typically, a text consists of a mixture of discourse within a particular domain and metadiscourse about it. SL involves:

- the sharing of common knowledge which includes the set of facts, assumptions and understanding underlying a text
- the restricted purpose of a text

This latter notion is close to 'utterance purpose', i.e. the overall reason for the utterance (Grosz, 1982:163) for discourse analysis. This factor is related to the specific need for a text, i.e. the communicative use of SL. Here we are also referring to the communicative functions of a text (Biber 1988:28) which situate a text for us. Traditionally, these functions are linked with a set of specific linguistic features: we will examine these features later. Situations are characterised by a set of parameters including communicative roles of the participants, the relations among the participants, social evaluation etc. (cf. Halliday, 1978; Brown and Fraser, 1979).

3 Characteristics

In order to better appreciate the nature of SL, we now propose to look at the various characteristics of SL in more detail. From these characteristics we will examine those which are relevant for MT.

Lehrberger (1982) gives the following characteristics of a SL:

i. limited subject matter

ii. lexical, syntactic and semantic restrictions

iii. 'deviant' rules of grammar

iv. high frequency of certain constructions

v. particular text structure

vi. use of special symbols

(i) By limited subject matter we mean texts with domain specific knowledge eg. immunology, computer maintenance etc. The immediate implication of the first characteristic is the use of 'special words'. SL is not characterised by lexical features alone. Vocabulary is an important factor but it is not sufficient. This is often forgotten (or more commonly ignored) by those coming from a non-sublanguage background, and is probably one of the reasons why sublanguage studies have not had any important status in Eurotra.

(ii) restrictions

a. lexical restrictions

Vocabulary restrictions characterize a SL. There are typically many technical terms that have precise meaning. E.g. 'floating point', 'eyebolts' (from our corpus of lathe manuals), 'peroxide' etc. Other words do not occur at all; in our corpora we did not find verbs such as 'to love', 'to hate' etc. We might find restrictive use of personal pronouns 'I', 'me', 'us' etc. The set of words which characterise different SLs are not mutually exclusive. There is also overlap (in terms of homonymy, for example) with GL. The following identifies just some ways of SL designation (terms) that borrow from GL:

- simile e.g. leg-type brush holder, tooth-shape
- metaphoric use of words from GL e.g. tooth of a gear, mouth of a furnace
- redefinition, reducing the extension of a GL word for the specific reference within a subject domain. (cf. Sager et al., 1980)

SLs have developed elaborate systems of concepts, i.e. their own special referential systems. The differentiation between general and special reference is expressed in lexical terms in a distinction between 'terms' which have special reference within a particular discipline, and words which function in general reference over a variety of subject fields. This characteristic is reflected in selectional restrictions among word classes, in SL co-occurrence constraints. Harris (1968) focused on the co-occurrence properties of words in SL. We will discuss the constraints further below when we consider semantics.

b. syntactic restrictions

Syntactic restrictions occur as a combination of SL and text type. Text typology will be examined below. Work in text typology (Biber, 1988) identified a set of linguistic features (not only syntactic) which are associated with particular communicative functions and text types. In SL literature (Sager, 1986; Kittredge 1987; Lehrberger, 1982 among others) scholars refer to weather bulletins, market reports, aircraft maintenance manuals, abstracts in the field of lipoprotein kinetics etc. The syntactic constraints met with in these texts were of the following type:

- sentence types: absence of direct questions, tag questions, certain types of nominalisations.
- tense and aspect: predominance of simple present; restrictive use of past tense; almost complete absence of future tense.
- modality: only certain types of verb modals used
- ellipsis: of articles, subject, copula
- long sequences of nouns:

stability augmentor pitch axis actuator housing support

(Lehrberger 1982:84)

Long sequences of nominals are very frequent in many SL texts, and pose problems in their correct bracketing and interpretation. We will return to this phenomenon below while examining the specific semantics of SL.

c. semantic restrictions

We saw that the restricted subject domain limits the vocabulary and the inventory of syntactic structures in the SL. More important than the limitation in the size of vocabulary are the SL co-occurrence constraints which may be viewed as a manifestation of the underlying semantic constraints

of the domain. Research (Hirschman, 1986) showed a correlation between SL classes obtained from co-occurrence phenomena and basic semantic categories perceived in a domain.

A major semantic property of SL is the reduction in polysemy. We distinguish the following cases:

(i) a word may occur only in one category in the SL

cable_N *but* *cable_V the forward compartment

(Lehrberger, 1982:85)

(ii) the range of meanings of a word within a given category is restricted

bore_N is a cylindrical hole or the inside diameter of cylinder

therefore: *the pilot is a bore

Reduction in the number of categories to which the individual words belong results in fewer combinations and less ambiguity.

$\text{check}_{N,V}\ \text{pump}_{N,V}\ \text{case}_{N,V}\ \text{drain}_{N,V}\ \text{fitting}_{N,V}$

Instead of having 32 paths to be explored the fact that we know that 'case' is not used as a verb in the corpus and is listed in the dictionary as a noun reduces the number of combinations to be tested.

The semantic restrictions of the domain are reflected in the kind and the number of semantic features for parsing. A noun which designates either concrete or abstract objects in the language as a whole could be used only as concrete in a SL, for example.

Also we can define restrictions in the kinds of subjects or objects SL verbs can take:

act: subject [-animate]

SL semantic patterns correlate with observed linguistic patterns and encode certain aspects of domain knowledge. The semantic patterns in a SL represent the knowledge expressed in domain-specific texts. They are typically represented by information schemata and can be considered frame representations of the domain. For more information we refer the reader to Marsh (1986), Hirschman & Sager (1982) and Minsky (1975).

(iii) 'deviant' grammar rules

By 'deviant' we mean with respect to the GL rules. Such rules are considered normal in a SL but are ungrammatical in GL. These rules also refer to specific co-occurrence restrictions met only in the SL.

If we examine the following sentences:

i. the patient presented with influenza

ii. the patient presented to the doctor with influenza

iii. *the patient presented the doctor with influenza

Sentences (i) and (ii) are grammatical within the SL of a medical doctor referring to a patient and (iii) is ungrammatical within the SL. On the other hand, the frame of the verb 'present' is ungrammatical with respect to GL in sentences (i) and (ii). For those who might remain unconvinced by this example, we point out that general lexicographers from Collins and Oxford University Press, when asked at this MT workshop to comment on the 3 sentences above, noted that they did not accept (i) and (ii) as 'English' i.e. general English and would not include such a reading for 'present' in their general dictionaries.

Harris (1968) mentions that the language of biochemistry accepts "The polypeptides were washed in hydrochloric acid" but excludes "Hydrochloric acid was washed in polypeptides", which is acceptable in GL.

By deviant we further mean that a SL has a grammar of its own which is not just a subset of the rules of the grammar of GL.

(iv) high frequency or low frequency of certain constructions

SL texts are characterised by the frequency of occurrence of some constructions such as:

- high frequency of imperatives e.g. 'check', 'add' in manuals
- long noun sequences are very frequent in some SLs

Nominal compounds are notoriously hard to analyse, at least in English, and available means are typically unable to provide good solutions. Their correct analysis requires extensive use of semantic relations. The fact that long nominal compounds are ambiguous (either due to the imprecision of available techniques or to natural ambiguity) does not limit their usage; on the contrary they are very frequent in the context of SL. Work on synthesizing semantic interpretations of compounds can be found in Finin (1980) and Finin (1986). In the LSP project long nominals were called 'empilages' (Lehrberger 1982:92). The proper bracketing of an empilage requires an understanding of the semantic/syntactic relations between the components. Assuming the nominal compound:

we must know that 'main' applies to 'fuel system' not to 'fuel' or 'drain valve'. Contextual and real-world knowledge is increasingly thought necessary to deal with such phenomena.

In order to obtain correct bracketing of N+N sequences it is crucial to recognise co-occurrence restrictions. Suppose a given noun 'n' can bear a certain relation to an immediately following noun. This does not mean that 'n' bears that relation to any noun that happens to occur immediately following it.

For example: 'installation' indicates FUNCTION in 'installation kit' and in 'installation procedure' but not in 'installation difficulty'.

There are many possible relationships between two nouns forming a compound, according to the different domain/context.

In the case of 'installation', the subclass of nouns to which this noun can bear the relation FUNCTION in the specific domain is specified (TAUM, Lehrberger 1982:103). This information can be made available by indicating in the dictionary entry of a noun the relations of this type in which the noun participates in the SL. This task is feasible in a restricted domain where the number of relations is not so great. Research at TAUM resulted in the definition of about 50 relations some of which were already proposed in the linguistic literature (Levi, 1978; Downing, 1977).

One could assign noun complementation in the dictionary to indicate the possible semanticosyntactic relations between noun compounds.

In addition, nominal compounds exhibit a variable degree of non compositionality. This notion is synonymous to idiomacity and results in lexicalisation. The same compound could be fully compositional in one domain and non compositional in another.

(v) particular text structure

Variation occurs across text types within the same SL. By text type we understand the text structure layout, the way the information is displayed, the style. Here we could redefine some terms. What we understand as SL is a combination of two parameters: the subject domain (sharing of lexical, semantic, pragmatic, syntactic restrictions to an extent) and the text type (immediately identifiable in syntax). We talk about manuals for aircraft maintenance or computers sharing the same syntactic properties such as the use of imperatives, the absence of some constructions, ellipsis etc. Things however are not so clear cut. The general characterisation of 'manual' is not enough. We need to define other factors that give rise to different text types. In the definition of SL in the beginning of this paper we referred to communicative functions that give rise to different SLs or text types.

The purpose, use of a text results in differences in text structure. Again the differences are mainly syntactic than semantic. Structural resemblances between semantically different SL (cooking recipes and aircraft manuals) are related to similarities of text purpose (cf. Kittredge, 1982).

The differences within some text type across subject domains have to be delineated. Some SLs exhibit more 'rigidity' of style than others, e.g. the guidelines for writing scientific prose are variously stricter or looser. Some SLs do not have strict norms for composition of texts to avoid ambiguity. Other areas to be examined are:

a) to what extent the same text type varies cross-linguistically

Research has been done in the past comparing technical manuals in aviation hydraulics and cooking recipes, regional weather synopses and stock market reports (Kittredge, 1982) in English and French. The results were very promising especially for our purposes (MT). The key features examined for establishing the characteristics of SL were: the textual organisation of these SLs, their specific word co-occurrences, sentence structures (e.g. length, patterns of subordination, use of passives) and use of tenses. As a result Kittredge was able to conclude that:

> *parallel sublanguages of English and French are much more similar structurally than are dissimilar sublanguages of the same language. Parallel sublanguages seem to correspond more closely when the domain of reference is a technical one.*
>
> (Kittredge, 1982:109)

There was identity of structures which were absent from both languages. In the case of parallel structures the same frequency of occurrence was identified:

> *each SL seems to move away from its respective language norm in the same way when cross-linguistic comparisons are made.*
>
> (Kittredge, 1982:129)

b) which linguistic features belong strictly to the SL and which belong to the text type?

It is apparent at this stage that more work needs to be done on a large scale to verify earlier findings and to identify the borderlines between SL and text type.

4 Benefits from adopting a sublanguage approach to MT

Given that SLs can be distinguished from instances of GL, the question arises for those interested in NLP:

> *how can we use what is known about the cohesive properties of individual SLs to improve techniques for parsing and synthesizing coherent texts?*
>
> (Kittredge, 1982:110)

The implications of the variation and homogeneity of SL for NLP are interesting. Since SL grammars differ within the same language (i.e. there are different structural types) separate grammars will be needed in order to analyse each text type (on the level of syntax and semantics). That is, apart from the obvious differences in vocabulary, a parser specific to a given SL would provide positive results in NLP, especially MT. By specific, we do not necessarily mean a dedicated system: modern NLP system design emphasizes modularity and flexibility, therefore we require e.g. of a parser that it be easily customizable, by e.g. changing the grammar which drives it, its dictionaries etc. Conceptually, we think of each sublanguage having its own description. This may be reflected in practice, especially in initial work, as it seems somewhat inefficient if not actually an immense hindrance to think in terms of some overall general description being specialised to some SL description. For example, it would seem a bad idea to take existing Eurotra descriptions and specialise them for one SL, then to take that result and incorporate filters and restrictions for another SL.

Sublanguage co-occurrence patterns help to reduce the syntactic ambiguity of the source language analysis. SL sentence types lead us to formulating target structures for semantic representation. A SL grammar may eliminate ambiguity in some structures drawn from the GL.

The lexical restrictions discussed in the characteristics of SL suggest an obvious advantage in adopting a SL approach while building our dictionaries. The specific complementation and co-occurrence restrictions can be coded enhancing correct and non-ambiguous parsing. This has also the advantage of reducing the size of dictionaries. The important point here is how indeed does one arrive at significant generalisations about co-occurrence behaviour within a SL? Below, we discuss automatic methods for arriving at such information (cf. section on clustering below).

We are all aware of TAUM METEO as an example of the utility and application of SL studies in the domain of MT. This SL has a very small vocabulary and is characterized by a telegraphic style. The syntax is also very restricted. TAUM demonstrated how a SL approach could be useful in MT. On the other hand, the highly restricted text type and SL of TAUM might drive some people into

thinking that this was an isolated example. This brings us to the issue of the 'representative corpus'. The idea of a representative corpus is related with the notion of closure of SLs. Eventually, it is stated in the literature on this topic, apparently divergent texts of some type will converge in terms of phenomena, if we are in some sublanguage. However, we should note that:

> *Estimating the computational tractability of sublanguage texts goes beyond the question of sublanguage closure. In the case of MT, the feasibility of correctly analyzing the source language texts is somewhat separate from the transfer problem.*
>
> (Kittredge, 1987:66)

The set of relevant questions to be asked to determine computational tractability concerns the set of linguistic phenomena we examined while establishing the characteristics of SL, e.g. ellipsis of articles, copula, use of long nominal compounds, co-referential links etc.

Syntactic rules in MT cannot adequately characterise phenomena such as:

- lexical ambiguities: polysemy, homonymy
- structural ambiguities: internal structure of compounds, PP attachment etc.
- textual links: anaphora, ellipsis etc.

Each of these phenomena may cause wrong translation. Semantic information may help resolve some of these ambiguities which is more feasible in a restricted domain e.g. establishing sets of semantic features, semantic relations for SL instead for GL.

In some cases the sole availability of sentential meaning representations is not sufficient to solve the above mentioned problems. Instead, the use of textual meaning representations which encompass properties such as: logical consistency, compatibility with domain-specific knowledge and with situation knowledge would represent a better alternative. Here, we step into Knowledge Representation techniques which point the way towards full semantic text representation. Research on text types, and in general on text linguistics and discourse are important for us also, as we need to be able to determine intersentential, inter-paragraph, etc., relationships in our effort to constrain possible interpretations even within some sublanguage.

One highly interesting possibility from the MT point of view in the way sublanguage studies have been oriented, recently, is that a certain amount of interlinguality may be utilised. It is typically stated that SL structures are presumed to be the information structures of the subject matter (cf. the work of Harris and the LSP team). These semantic patterns are not an interlingua themselves

but provide a framework for meaning representation and are a step closer to an interlingua. In our case they would result in simpler transfer, as the notion of parallel SLs in different natural languages exhibiting more structural similarities than different SLs within the same natural language clearly indicates.

BIBLIOGRAPHY

Ananiadou, S. (forthcoming) Automatic Term Recognition, Edinburgh University Press.

Biber, D. (1988) Variation across Speech and Writing. Cambridge University Press.

Downing, P. (1977) "On the Creation and Use of English Compound Nouns", Language 53 (4).

Finin, T. (1986) "Constraining the Interpretation of Nominal Compounds in a Limited Context", in Grishman & Kittredge (eds) pp. 163-173.

Finin, T. (1980) "The Semantic Interpretation of Compound Nominals", Proceedings of the First National Conference on AI, Stanford.

Harris, Z. (1986) Mathematical Structures of Language. Wiley-Interscience.

Harris, Z. et al. (1989) The Form of Information in Science, Kluwer Academic Publ.

Hirschman, L. (1986) "Discovering Sublanguage Patterns", in Grishman & Kittredge, pp. 211-234.

Grishman, R. & Kittredge, R. (1986) Analyzing Language in Restricted Domains, Lawrence Erlbaum Ass.

Grosz, B. (1982) "Discourse Analysis" in Kittredge R. & Lehrberger, J. (eds) pp. 138-174.

Lehrberger, J. (1982) "Automatic Translation and the Concept of Sublanguage", in Kittredge & Lehrberger, pp. 81-106.

Kittredge, R. (1982) "Variation and Homogeneity in Sublanguages", pp. 107-137.

Kittredge, R. & Lehrberger, J. (eds) (1982) Sublanguage. Studies of Language in Restricted Semantic Domain, Walter de Gruyter.

Kittredge, R. (1987) "The Significance of Sublanguage for Automatic Translation", in Machine Translation, S. Nirenburg (ed.), Cambridge University Press, pp.59-67.

Kosaka, M., Teller, V. & Grishman, R., (1988) "A Sublanguage Approach to Japanese-English Machine Translation", in New Directions in MT, Maxwell D. et al. (eds). Foris, pp. 109-120.

Marsh, E. (1986) "General Semantic Patterns in Different Sublanguages", in Grishman & Kittredge, pp. 103-127.

Minsky, M. (1975) "A Framework for Representing Knowledge", in P. Winston (Ed.), The Psychology of Computer Vision, McGraw-Hill.