SUBLANGUAGE, TEXT TYPE AND MACHINE TRANSLATION

Sharon O'Brien, B.A.

A thesis submitted to Dublin City University in candidacy for the degree of Master of Arts

SEPTEMBER, 1993

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Arts is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ **Date:** _____

Supervisor: Dr. Jenny Williams
School of Applied Languages
Dublin City University

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank every individual who helped me in any way to write this thesis. Small and big favours alike were greatly appreciated. A very warm "thank-you" in particular go to:

- **Dr. Jenny Williams, School of Applied Languages, D.C.U.**
- **Professor Juan Sager, UMIST**

- **Jennifer Pearson, Elaine Quinlan, Donncha Ó Croinín, and Dr. George Talbot at the National Centre for Language Technology, D.C.U.**

- **The Eurotra Scholarship Programme which provided funding for the research.**

- **My Family, for their never-ending love and support.**

- **Ruairí, do gach rud!**

- **Annette, for patient proof-reading and friendship.**

- **Muintir Uí Ailín, for encouragement and support.**

- **Stiofán and Stephen, for help with printing.**

- **And God.**

Go raibh míle maith agaibh go léir!

ABSTRACT

This thesis explores the domains of sublanguage, machine translation and textual analysis. Chapter 1 discusses the definitions and characteristics of sublanguage put forward by researchers to date, as well as the background of textual analysis in linguistics. This discussion reveals that, although there is much to be gained from textual analysis, little consideration has been given to the notion of "text" in the sublanguage approach to machine translation (MT).

Before any sublanguage analysis can proceed, compilation of a corpus is necessary. To date, attention has been focussed on the criteria for compiling general language corpora. Chapter 2 addresses the problems of compiling corpora for sublanguage research and offers guidelines for this purpose.

An exploration of the advantages of considering text type and communicative function in the sublanguage approach to MT is the focus of Chapter 3. Three text types with a similar communicative function from the same highly restricted sublanguage domain are compared for linguistic features which cause semantic, syntactic and lexical ambiguities.

Finally, Chapter 4 summarises and evaluates the results obtained in Chapter 3. Conclusions are drawn about "text type" and communicative function and about the advantages of considering "text" for MT.

LIST OF ABBREVIATIONS

ADJ	-	**Adjective**
ADV	-	**Adverb**
AI	-	**Artifical Intelligence**
FAHQ(M)T	-	**Fully Automatic High Quality (Machine) Translation**
LSP	-	**Language for Special Purposes**
MAT	-	**Machine Aided Translation**
MOD	-	**Modifier**
MT	-	**Machine Translation**
N	-	**Noun**
NLP	-	**Natural Language Processing**
N (Mod)	-	**Modifying Noun**
NP	-	**Noun Phrase**
SL	-	**Source Language**
TGP	-	**Text(s) for General Purposes**
TSP	-	**Text(s) for Special Purposes**
TL	-	**Target Language**
V	-	**Verb**
V (Mod)	-	**Modifying Verb**
VP	-	**Verb Phrase**

CHAPTER 1 :

INTRODUCTION

1.0 The Definition of Sublanguage

One of the first uses of the term "sublanguage" was by Zellig Harris in 1968. He commented on the idea of a sublanguage being a subset of natural language and likened sublanguage to the notion of a subsystem in mathematics. Harris states that

> Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it.[1]

From this statement the notion of closure was derived which is one of the main properties of any sublanguage.[2]

During the 1960's, interest in the notion of sublanguage increased in tandem with an interest in modelling language using computers and in translating from one language into another with the aid of computers. Indeed, the revelation that language for general purposes was too ambiguous for "Fully Automatic High Quality Machine Translation" (FAHQMT) increased interest in restricted language, or sublanguage, even further. It was hoped that although general language could not be translated automatically by computer (at least not to the extent that it produced high quality output without human aid), then at least automatic translation of restricted language was a realistic aim.

Given that this is the background in which sublanguage research was fostered, it is not surprising that the focus of research has been on developing *applications* for sublanguage rather than on discussing the theoretical foundations of the concept. This is not to suggest that sublanguage researchers have neglected to define the concept of sublanguage, rather they have been so preoccupied with developing useful applications that they have not yet agreed on a standard definition for the term. When we consider

---

[1] Harris, Z., *Mathematical Structures of Language*, (New York: Wiley & Sons, 1968), p.152.

[2] The notion of "closure" and other sublanguage characteristics will be discussed further in section 1.2, Chapter 1.

the type of sublanguage projects carried out since the 1960s, it is obvious that the focus was on developing NLP applications rather than on theoretical discussion.

For example, one of the first investigations into the structure of a science sublanguage was undertaken in 1969 under the auspices of the "National Library of Medicine" in America. Textual material in the domain of pharmacology was examined and it was shown that it is possible to write a sublanguage-specific grammar for texts in a narrow science sub-field. This study also proposed the concept of an information format as a means for representing the information in sublanguage texts. The information format organises the sublanguage sentence types into a tabular representation so that the document can be quickly inspected.

Following on from this study, Naomi Sager and a team of researchers at New York University developed an automatic parser for medical language texts as part of the NYU Linguistic String Project in 1981. On the basis of their success in this project the researchers concluded that it was possible to refine these techniques to create a powerful, general system for large scale automatic processing of natural language documents in a given sublanguage.[3]

In 1977, a broad study of English and French sublanguages was initiated within the "Contrastive Syntax Project" of the Université de Montréal. This project addressed eleven varieties of sublanguage which included macro-economics, children's stories, literary criticism, weather bulletins, aviation hydraulics manuals, pharmacology reports, weather synopses, recipes, stock market reports, micro-economics (mathematical foundations) and university catalogues (degree requirements). The first three were subsequently declared too diverse to qualify as sublanguages. During this project some interesting questions about the nature of sublanguages were addressed, for example,

---

[3]Hirschmann, L. & N.Sager, "Automatic Information Formatting of a Medical Sublanguage", in: Kittredge, R. & J. Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, (Berlin & New York: de Gruyter, 1982), pp.27-80 (p. 67).

- Do parallel sublanguages of different languages show resemblances due to their shared semantic and pragmatic conditions, or is it due to stylistic borrowing between technical subcultures in contact?

- How does the cohesiveness of text vary from sublanguage to sublanguage?

- How do the constraints of sublanguage semantics and pragmatics influence sentence and text structure?

Some of these questions have not been answered to date.

Conclusions from this project indicated that parallel sublanguages in French and English are much more similar structurally than non-parallel sublanguages of the same language. Furthermore, if the domain is technical, the correspondence between the languages is greater than when the domain is non-technical.

At the same time the team in Montréal were working on producing the first sublanguage-based machine translation system, TAUM METEO, which translates weather reports from English into French with a high percentage of accuracy. Today, Taum Méteo is regarded as one of the most successful MT projects ever undertaken. Following that, the Montréal team turned their attention to the analysis of aviation hydraulics manuals which they found to be more complex than the sublanguage of weather reports.[4]

[4]Kittredge, R. & J. Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, (Berlin & New York: de Gruyter, 1982), Chapters 2 & 3.

More recently, the European-wide MT research project, *Eurotra*, also recognised that the sublanguage approach presented the most hopeful path for successful machine translation and research on the sublanguage of satellite telecommunications was initiated during that project.[5]

As already suggested, the main aim of many of these pioneering projects was to describe the grammar of particular sublanguages and to develop natural language processing applications for them, rather than to define the notion of "sublanguage" in general. Kittredge, who was involved in the Contrastive Syntax Project at the University of Montréal, comments on the lack of an empirically adequate definition of the term sublanguage.[6] According to him, there is still a need to develop some criteria for deciding what the limits are for a given sublanguage, and whether closely related varieties of language should be considered parts of the same sublanguage or as constituting separate systems. For him, Harris' property of closure, which states that a sublanguage has a finite set of grammatical constructions, is not in itself sufficient to resolve these questions:

> The term *sublanguage* has come to be used not just for any marked subset of sentences which satisfies the closure property, but for those sets of sentences whose lexical and grammatical restrictions reflect the restricted sets of objects and relations found in a given domain of discourse.[7]

Hirschmann and Sager's definition of sublanguage focuses not only on the restricted nature of the language used but also on the users/producers of the language. They define sublanguage as:

> The particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles on a technical speciality or science subfield), in which the authors of the documents share a common vocabulary and common habits of word usage.[8]

---

[5]Indeed, many of the ideas contained in this thesis were first developed when working with the Eurotra-Ireland team on the "Sublanguage and Text Type" project.

[6]Kittredge, R., "Variation and Homogeneity of Sublanguages", in Kittredge & Lehrberger (eds.), op cit, p.110.

[7]Kittredge, R., in Kittredge & Lehrberger (eds.), op cit, p.2.

[8]Hirschmann, L. & N. Sager, op cit, p. 27.

It follows from this definition that a new term or grammatical construction does not become a true part of the sublanguage until its use has been conventionalised by the community of speakers.

Lehrberger describes sublanguage more generally as "an independent system or as a subsystem of the natural language".[9] His outline of sublanguage characteristics must be seen in the context of his discussion on the place of sublanguage in relation to the language system. He believes that one way of looking at the structure of sublanguage is to look at the structure of natural language, of which it is a part. Once again, the analogy of a subsystem in mathematics is used to describe a sublanguage of language. Lehrberger, however, does not believe that the link between sublanguage and language has been clearly described and therefore explores the notion of how the two are linked.

The initial question to be addressed here is: what is the definition of "language"? In much of the literature on sublanguage theory and analysis, "language" is often referred to as *natural language, standard language, general language* or *language as a whole*. Lehrberger points out that when "natural language" is being discussed it is not made explicit whether it is taken to mean the full range of lexical, syntactic and semantic features inherent in a language, or whether it means the percentage of the full inventory of language as spoken by a community of speakers. He also makes it clear that a natural language is not merely the composition of all possible sublanguages. It is not known how many sublanguages exist in a given natural language, and this would be difficult to ascertain because sublanguages come into existence over a period of time through the use of language in specialised domains by experts in those domains, and

[9]Lehrberger, J., "Sublanguage Analysis", in Grishman & Kittredge (eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, (Hillsdale, New Jersey & London: Lawrence Erlbaum Associates, 1986), pp.19-38 (p.19).

also the boundaries of sublanguages are sometimes obscure since many of them overlap and, therefore, their interrelations are difficult to measure.

Assuming that the term "language as a whole" covers all subject-specific language and general language texts, it would seem that a grammar of the language as a whole should describe all the sublanguage grammars in it. Despite the fact that a sublanguage is recognised as being a part of natural language, Lehrberger points out that there is no evidence to support the view that the grammar of a sublanguage of, for example English, must be a subgrammar of the grammar of English as a whole as illustrated in the following sentence:

(5) Refill tank if fluid level low

This sentence is accepted as being grammatical in a sublanguage of English but may not be regarded as grammatical in English, and yet is acknowledged as being a sentence of English. Lehrberger clarifies the concept of sublanguage grammar by distinguishing between discourse and sublanguage. If L' represents a sublanguage with a semantic domain D, the sentences of L' are those that express the properties and relations of D. Actual discourse is also used in L' and it offers an insight into the world represented by such a sublanguage. However, only a percentage of the material in the discourse expresses the properties and relations of D. Therefore, he concludes, when writing a grammar for NLP, actual discourse must also be processed which means that the sublanguage grammar has to be even more comprehensive than a grammar of the theoretical construct L'.

If L is then taken to be the whole language which includes all varieties of spoken and written language and an indefinite number of sublanguages, Lehrberger speculates as to how a grammar of L would appear and how a sublanguage would be formally described as a subset of L. He considers it unlikely that a grammar of L would be

obtainable. However, a grammar of standard language, *Lstd*, may be obtainable. This leads him to put forward a "paraphrase relation" which states that "whatever can be said in a sublanguage of a natural language *L*, can be paraphrased in *Lstd*".[10]

If *L* is language as a whole, *L'* is a specific sublanguage and *Lstd* is the standard language, then the relation between them is such that *L'* intersects *Lstd* in *L*, as can be seen in Fig.1.1. A sublanguage *L'*, of a natural language *L*, can be viewed as resulting from restrictions on and deviations from the grammar of *Lstd*.

[10] Lehrberger, J., in Grishman & Kittredge (eds.), op cit. p.20.

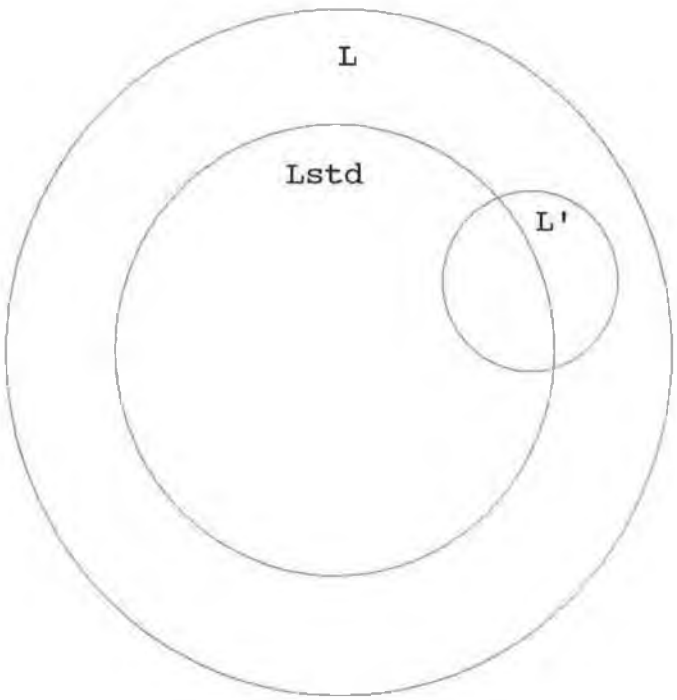

Fig. 1

Lehrberger stresses, however, that the approach taken to account for the position of sublanguage in language depends on the aim of the research. For example, some sublanguages of a natural language could be treated as independent systems rather than situating them in *L* and relating them to *Lstd*, an approach which may be beneficial for some practical applications restricted to particular domains. Another option would be to describe the sublanguage in terms of the structures and operations of the larger systems. One thing, however, is clear. While sublanguages are dependent on general language for their existence, they perform much more specialised tasks:

> The sublanguage 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.[11]

Lehrberger, like Kittredge, recognises that the problem of writing formal grammars for natural languages aimed at MT may be resolved if a subset of natural language is used, as it is easier to write a set of rules that will generate all the possible sentences of a special-subject field, rather than produce rules that will generate all possible sentences for a natural language.

Hoffmann is one of the pioneers of research on restricted language (or *Fachsprache*) in the German school. He dates the study of restricted language back to the 1930s but states that the concept really only came to the fore during the mid-1970s. His definition of *Fachsprache* is similar to Sager and Hirschmann's definition in that it considers specialised language to be language used in a restricted domain for communication between experts:

> Fachsprache - das ist die Gesamtheit aller sprachlichen Mittel, die in einem fachlich begrenzbaren Kommunikationsbereich verwendet werden, um die Verständigung zwischen den in diesem Bereich tätigen Menschen zu gewährleisten.[12]

Hoffmann identifies seven areas which comprise LSP research:

(i) Terminology

(ii) Functional Speech Analysis

(iii) Functional Stylistics

(iv) Language of Commerce (Wirtschaftslinguistik)

(v) Science and Philosophy (effects of LSP on the language system as a whole)

(vi) LSP for translation purposes

---

[11] Lehrberger, J., in Grishman & Kittredge (eds.), op cit, p.22.

[12] Hoffmann, L., *Kommunikationsmittel Fachsprache*, (Berlin: Akademie-Verlag, 1985), p.53.

(vii) Teaching of Sublanguages (*Subsprachen*)

The boundaries of each of these areas are not fixed with the result that there is frequent overlapping. Thus, in translation studies, for example, one might encounter an analysis of specialised terminology.

Hoffmann uses both the term *Fachsprache* and the term *Subsprache* but the distinction between the two is not clear. From the above classification, it would seem that *Fachsprache* is a more general term which refers to language used for special purposes (or LSP in English) or any branch of linguistics associated with the study of restricted languages. *Subsprache*, on the other hand, refers to a *specific* LSP from a particular domain, which is taught to non-native speakers of the (sub)language, e.g. Business English, German for computer science students, and so on (i.e. number (vii) in Hoffmann's list above). This does not correspond to the usage of the English terms *LSP* and *sublanguage* in English since, today, the latter is generally used in computational circles only while the former is generally used in applied linguistic circles to refer to the teaching of a specialised language to non-native speakers of the language. To complicate matters further, Hoffmann later refers to the *Fachsprache* Physics as an object of a more general *Subsprache* which, in turn, is part of the *Nationalsprache (Gesamtsprache)*, or the whole language system[13] Here the term *Fachsprache* is being used to denote a particular sublanguage (i.e. Physics) which forms part of a more general, yet still specialised, *Subsprache*. One thing is clear, Hoffmann does not use the term *Subsprache* to refer to the analysis of restricted languages for computational purposes. Despite the confusion and lack of symmetry between terms, his approach to the analysis of LSP texts could prove very useful for computational linguistics. He proposes eight stages of analysis:

[13]Hoffmann, L., op cit, p. 51.

(i) Terminology and vocabulary research

(ii) Investigation into word formation

(iii) Semantic analysis of terms

(iv) Morphological research

(v) Co-occurrence analysis

(vi) Investigation of phrase structure

(vii) Investigation of sentence structure

(viii) Text and text type research

Hoffmann is one of the few linguists who recommends broadening the objectives of LSP research to take account of the text.[14]

He proposes a detailed classification system for LSPs which allows for a horizontal classification according to subject domain, e.g. medicine, chemistry, physics, mathematics etc., and a four-way vertical classification which has to do with (a) the degree of abstraction (ranging from very high to very low on a scale of five), (b) the amount of artificial language/specialised terminology contained in the sublanguage, (c) the environment in which it occurs (e.g. theoretical, applied, experimental etc.) and (d) the relationship between participants (e.g. specialist to specialist, specialist to apprentice and so on).

Like Lehrberger, Hoffmann seeks to understand the relationship between specialised language and what is commonly known as *general language*. There are several ways of looking at the problem. He outlines two possible solutions. On the one hand, general language is subsumed by the more superordinate concept of the whole language (or *Gesamtsprache*). Restricted languages are then part of - and subordinate to - *Gesamtsprache*. On the other hand, general language (*Gemeinsprache*) can be seen as

[14]This issue will be dealt with in more detail in section 1.4, Chapter 1.

the *standard* language of a particular speech community, which is a restricted language in its own right. It is classified as a sublanguage along with physics, mathematics etc. and is subsumed by the notion of *Gesamtsprache*. Having proposed these models, he then points out two shortcomings which would apply to any model, namely (a) it is difficult to define general language (or *Gemeinsprache*) as a sublanguage (*Fachsprache*) because, unlike most sublanguages, there is no fixed communicative situation in which general language occurs and (b) it is impossible to define the concept *general language* because every individual's knowledge of language is different. Also, language is constantly changing with new words being added and accepted by users and other words becoming obsolete. Moreover, words are frequently borrowed by sublanguages from "general language" to describe new concepts where they eventually take on the status of a term which forms part of the semantic network of the sublanguage domain. Therefore, it would be impossible to establish a fixed inventory of *general* vocabulary which could be used to differentiate general texts from specialised texts.

When Hoffmann's model describing the relationship between specialised language and general language is compared with that of Lehrberger, we find the latter more convincing. Hoffmann places each subject-specific sublanguage into distinct, separate boxes and describes them as being subordinate to language as a whole (which subsumes the notion of general/standard language). Lehrberger's model is more dynamic. He sees sublanguages as *intersecting* both standard language and language as a whole (See Fig. 1). The weakness in Lehrberger's model is that there is no consideration given to the factors which might influence the extent to which a sublanguage intersects with standard language and language as a whole such as text type, communicative function, degree of abstraction, relationship between participants etc. This is where Hoffmann's criteria for the vertical classification of sublanguages, which are listed above, could be very useful.

Sager et al also consider the relationship between sublanguages (or "special languages") and general language.[15] According to them, all special languages have dialectal and sociolectal variants. For this reason, they are often described as sub-languages of national languages or varieties of standard languages. This is incorrect in the authors' view because special languages often become supranational. In their opinion, special languages can only be contrasted meaningfully with general language which they define as

> an abstraction derived from a society's division of knowledge into general and special.[16]

For them, special languages are intersecting subsystems which overlap general language and are dependent on it. Most messages have some special language features and it is only when a certain density is reached that the message is described as "special". The measure of this density is decided separately by each society. Following on from this, a special language without any general language elements is an artificial language, in their opinion, because it loses its ability to be its own metalanguage.

Like Hoffmann, Sager et al question the notion of the term "general language". They suggest that it is determined separately by individual communities:

> There are as many general languages as there are special languages because each group of specialists has a different notion of what constitutes the general knowledge which forms the basis of their subject. General knowledge also varies with the level of education of specialists and this in turn determines what is considered general language.[17]

Their definition of special language is based on their assumptions about its relationship to general language:

---

[15]Sager, J.C. et al, *English Special Languages: Principles and Practice in Science and Technology*, (Wiesbaden: Brandstetter Verlag, 1980).

[16]Sager et al, op cit, p. 64.

[17]Sager et al, op cit, p. 64.

> Special languages are semi-autonomous, complex semiotic systems based on and derived from general language...their use presupposes special education and is restricted to communication among specialists in the same or closely related fields.[18]

Again, like Hoffmann, Sager et al are among the few sublanguage researchers who have accounted for the text in their theory of sublanguage. They discuss the pragmatic features of the "special text unit" which is characterised by a unity of topic, reference and syntactic cohesion and by a traditional form which is determined by the author's intentions.

1.1 Summary

The foregoing discussion presents an overview of definitions which have been formulated to define the term "sublanguage". Some definitions describe the linguistic features of sublanguages, others focus on the users and producers of the sublanguage and others still focus on the relationship between general language and sublanguage. Although each of these definitions touch on some important aspect of sublanguage, no one definition fully explains the meaning of the concept "sublanguage". In order to understand the concept more fully, we must examine the characteristics of sublanguages more closely.

1.2 The Characteristics of a Sublanguage

Nobody would disagree with the statement that the language used in certain specialised domains is different from the language used for day to day communication. If we were asked, however, to explain how they differ, we would be faced with several problems.

---

[18] Sager et al, op cit. p.68-69.

Some of these problems have already been discussed, for example, the fact that there is no agreed definition for sublanguage, or that the notion of general language, against which sublanguage is frequently compared, is extremely difficult to pin down. A third problem is that there are no fixed criteria for identifying a sublanguage. Harris isolated *closure* as one of the main characteristics. Kittredge proposes "shared habits of word usage" by the speakers of the sublanguage as another property[19] Defining the community of speakers of a sublanguage, though, is not an easy task especially in written communication where there is free access to texts. Kittredge suggests that in cases such as these, different sub-types evolve which are directed at different types of user with varying levels of expertise. Linguistic characteristics vary from sub-type to sub-type and, according to Kittredge, it is reasonable to call each sub-type a sublanguage for computational and linguistic taxonomy purposes. It should also be noted that although sublanguages are for communication between experts in a community, non-experts are capable of understanding a certain percentage of what is said depending on how specialised the communication is. Therefore, "shared habits of word usage" by a community of speakers is inadequate as a defining property on its own.

On the basis of a contrastive analysis of a number of sublanguages, Kittredge identifies further sublanguage characteristics:

A sublanguage normally has a restricted lexicon. Precise measurement of the lexicon is difficult and estimates of lexicon size can be misleading. For example, in the sublanguage of recipes the lexicon seems rather extensive when all instruments and ingredients are considered. Nevertheless, Kittredge maintains that actual complexity is less than the lexicon suggests because the lexical items fall into a few specific categories. It is these categories that are important, not the number of lexical items:

[19]This property is also proposed by Hirschmann and Sager in their definition of sublanguage which was presented earlier.

> What counts most is the number of lexical categories and subcategories which must be distinguished for the proper grammatical description, and the average complexity of description of a lexical item in terms of these categories.[20]

On the level of sentence structure, Kittredge maintains that some sublanguage sentence structures cannot be easily compared with standard language sentence structures. This statement is not valid for all sublanguages, though, as other sublanguages show enough similarities that standard language can be used as a basis for comparison.

On the subject of linking devices in texts, Kittredge discovered that patterns are similar across identical sublanguages in different languages while they differ between different sublanguages in the same language. Moreover, technical sublanguage texts show stronger parallels than descriptive texts because the text purpose is more rigidly defined in the former.

Lehrberger identifies six characteristics by which a sublanguage may be identified. They are

(i) limited subject matter

(ii) lexical, syntactic and semantic restrictions

(iii) "deviant" rules of grammar

(iv) high frequency of certain constructions

(v) text structure

(vi) use of special symbols

We will consider each of these characteristics individually below.

[20] Kittredge, R., in Kittredge & Lehrberger (eds.), op cit, p. 125.

1.2.1 Limited subject matter

Sublanguages develop out of a need for communication between experts in a specialised field. Consequently, sublanguages are capable of describing a limited semantic domain only. Lehrberger maintains that restrictions on the semantic domain affect word usage, the size of the lexicon, and the inventory of semantic and syntactic features. This is very beneficial from the computational point of view because it reduces the amount of homonymy and polysemy, which in turn reduces the amount of ambiguity in the text. It must be noted, however, that even though a sublanguage may be labelled as belonging to a particular semantic domain, such as satellite telecommunications, it can also overlap with other semantic domains, such as economics, for example.

1.2.2 Lexical, syntactic and semantic restrictions

(i) Lexical Restrictions

Sublanguages are frequently compared (both with each other and with general language) in terms of the size and complexity of the lexicon. This type of comparison can be misleading, however, because a precise measurement is only possible if the sublanguage is "lexically closed", i.e. if no new lexical items can be added to its dictionary. It has already been pointed out (cf. Hirschmann and Sager's definition of sublanguage) that in dynamic sublanguages new words occur all the time. Thus, caution must be taken when comparing sublanguages on the basis of the size of their lexicon.

In general, a large portion of the natural language vocabulary will not appear in a sublanguage text and the number of lexical items in the sublanguage text will most

likely be smaller than in a general language text. Certain lexical items will be repeated with a high degree of frequency throughout the sublanguage text. Although repetition of lexical items is also characteristic of general language texts and is frequently used as a cohesion-producing device, the frequency of recurrence of lexical items in sublanguage texts is often a great deal higher than in general language texts. In the latter, the tendency is to replace lexical items with referential pronouns or with synonyms. The use of referential pronouns or synonyms, however, can often lead to ambiguity, something which must be avoided in special language texts, especially when the function is to instruct the user to carry out an action. Sometimes, lexical items that occur in sublanguages are not mutually exclusive, as some lexical items may be found in a number of sublanguages. The example that Lehrberger cites is the word "filter" which belongs to the lexicon of pharmacology, traffic signalling, electronics, and photography, to name but a few.

(ii) Syntactic Restrictions

Restricted syntax is another defining feature of a sublanguage. In most sublanguages the range of syntactic constructions is restricted as is syntactic creativity, e.g. the type of creativity found in poetic texts is not to be found in sublanguage texts. Lehrberger uses the language of aircraft maintenance manuals to illustrate how there are never any occurrences of direct or tag questions, simple past tense or exclamatory phrases in the corpus of texts. This facilitates automatic parsing but the benefits are cancelled out by the fact that sublanguage texts, including aircraft maintenance manuals, often have very short sentences with many conjunctions which cause parsing problems for automatic analysis, as in, for example,

(6) Disconnect pressure and return lines from pump

where it is not clear whether *return* is an imperative verb or a modifier of the noun *lines.*

(iii) Semantic Restrictions

The fact that a section of natural language may be restricted semantically means that the semantic features required for automatic language processing will be reduced in number and type. Lehrberger illustrates this by referring to the sublanguage of aviation hydraulics where many nouns which designate either concrete or abstract objects in language as a whole are used only concretely in the sublanguage, e.g. *spring, web, race*. Moreover, some words which designate both human and non-human in language as a whole are restricted to non-human in the above mentioned sublanguage, e.g. *agent, body, elbow*.[21] Finally, words that can occur in more than one grammatical category in language as a whole sometimes occur in only one category in a sublanguage.

1.2.3 Deviant rules of grammar

"Deviant rules of grammar" are what Lehrberger terms rules which are deemed ungrammatical in standard language, yet are accepted as normal in the sublanguage. One frequently recurring example is deletion of the article as in the following example:

Cap all open lines and ports

This example is taken from the sublanguage of aviation hydraulics manuals. The function of such manuals is to instruct and deletion of the article is one of the characteristics associated with that communicative function.

[21] Lehrberger, J. in Kittredge & Lehrberger (eds.), op cit, p. 87.

1.2.4 High frequency of certain constructions

It has already been mentioned that, in comparison with standard language texts, there may be a higher recurrence of lexical items in some sublanguage texts. Moreover, there may be noticeable occurrences of certain grammatical constructions which go beyond the normal frequency of occurrence in language as a whole, for example, on analysing our corpus of sewing instructions it was estimated that 68% of the total number of verb forms were imperative. Another example is the use of long sequences of NPs in technical manuals as exemplified below:

(8) external hydraulic power ground test quick-disconnect fittings[22]

1.2.5 Text Structure

It is not clear what Lehrberger means exactly when he lists "text structure" as a characteristic of sublanguages. Presumably he is referring here to the fact that most specialised texts have rigidly defined structures. Indeed, authors of some sublanguage texts are given guidelines to which they must adhere when producing manuals etc. However, it must also be pointed out that *all* texts have a structure. Not enough is known about the rules which govern the macrostructure of texts to claim that sublanguage texts are more structured than standard language texts. The issues of text, text structure and communicative function have largely been neglected in sublanguage studies. In the next section (1.4) we will examine this issue more closely.

[22]Lehrberger, J., in Kittredge & Lehrberger (eds.), op cit, p. 92.

1.2.6 Use of special symbols

Many sublanguage texts contain frequent use of special symbols, diagrams, graphs, acronyms, and abbreviations, e.g. *MHz, 4.8Kbits/sec, CVSD-PSK,* etc. The occurrence of these symbols is partly determined by the sublanguage (as in the above examples) and partly by the text-type (for example, the classification or numbering system for the different sections and subsections in a text is often determined by the text type and is therefore independent of the sublanguage).

1.3 Summary

In the above discussion we have outlined the general definition of a sublanguage, as proposed by various sublanguage researchers, and we have listed the characteristics by which a sublanguage may be identified. We have also discussed the debate about the status of sublanguage in relation to the whole language system.

Although there is no standard, accepted definition of the term *sublanguage* it is possible, on the basis of the above examination, to summarise the essential characteristics of the concept "sublanguage":

(i) A sublanguage is a subset of natural language which intersects the standard language in natural language. A sublanguage grammar results from restrictions on and deviations from the grammar of the standard language.

(ii) A sublanguage has lexical, syntactic and semantic restrictions when compared with standard language.

(iii) A sublanguage exhibits the property of closure. In other words, it has a finite set of grammatical constructions and semantic possibilities. Some sublanguages are still evolving. This means that new lexical items may be added to describe new concepts. These new lexical items usually fall into a pre-determined category or sub-category of the sublanguage grammar, thereby ensuring that the grammar remains "closed".

(iv) The restrictions on a sublanguage reflect the restricted set of objects and relations in the domain of discourse which that sublanguage seeks to describe.

(v) A sublanguage is used for communication between experts in a sub-domain. This does not imply, however, that non-experts are unable to comprehend anything that is being said in a sublanguage text.

(vi) In order to describe new concepts, a sublanguage borrows words from standard language to describe new concepts in the sublanguage which subsequently become part of the sublanguage lexicon or, if the sublanguage is being monitored by a terminology standardisation authority, the borrowed words may be replaced by a standardised term.[23] The extent to which sublanguages borrow from standard language varies from sublanguage to sublanguage and is dependent on the rate of development of a sublanguage (and the resulting requirement for new terms to describe new concepts).

(vii) A sublanguage has a high frequency of certain grammatical constructions and may use special symbols.

---

[23]It is also possible for new terms to be *coined* in sublanguages using, for example, acronyms such as "WYSIWYG" (What You See Is What You Get) from the sublanguage of information technology.

Finally, two sublanguage characteristics that have been noted but essentially neglected by sublanguage researchers are:

(viii) A sublanguage may have different text types with varying communicative functions. Within one sublanguage text, there may be more than one text type.

(ix) A sublanguage may be either oral or written in nature or both.

1.4 Text Type, Sublanguage and Machine Translation

1.4.1 Introduction

In the previous section, it was noted that Lehrberger lists "text structure" as one of the main characteristics of a sublanguage. Under this heading he considers "gross structure" of a text and "linking devices". By gross structure he means the division of a text into numbered sections. He notes that a word in one section may have a different translation when compared with the same word in another section.[24] The example he uses is the word *valve* from the sublanguage of aviation hydraulics which can be translated into either *soupape* or *valve* in French depending on which text section it appears in.

Under the heading of linking devices, Lehrberger considers devices which produce textual cohesion such as repetition of lexical items, use of referential pronouns, repetition with a change in grammatical category and so on.

However, little consideration has been given to the notion of text type/communicative function in sublanguage studies in general. Instead, sublanguage researchers have concentrated on providing descriptions of sublanguage grammars which are based on the syntax of *sentences* and/or *phrases,* but not on texts.

Studies of *text* in linguistics in general have provided us with invaluable information about cognitive models of text production and reception, about links between communicative function and the linguistic features of a text, and about methods of

[24]Lehrberger, J., in Kittredge and Lehrberger (eds.), op cit, p.95.

classifying texts into specific categories. Why, then, has sublanguage analysis and, more specifically, sublanguage analysis for machine translation, not taken text into consideration? Also, what are the advantages for MT studies in considering text? These are some of the questions which we will try to answer in the forthcoming section. Before that, however, it is necessary to map the development of textual analysis and to consider some of the main arguments and theories put forward in this discipline. Firstly, the development of text linguistics in general will be considered. This will be followed by a consideration of the work done in the area of LSP textual analysis as well as analysis of texts for translation purposes. Finally, efforts made by some researchers to incorporate a text-based approach into MT will be discussed and some recommendations, based on the conclusions of text analyses to date, will be made for a text-based/sublanguage approach to MT.

1.4.2 Development of Text Studies

At the same time as the sublanguage approach to machine translation first became popular during the 1960s, linguists who had previously considered the sentence to be the largest possible linguistic unit became aware of the limitations imposed by this belief and turned to the concept of *text* which they hoped would provide more insight into the workings of the language system. Since then, two main trends have emerged both in the area of textual analysis and restricted language analysis. The first is a theory-based approach to the analysis of texts, now commonly called "text linguistics". The aim of this approach is to describe text grammars and cognitive models for the production and reception of texts. This approach also includes analysis of texts for specific purposes, i.e. restricted language texts, with the aim of describing the syntactic and semantic features of those texts. The second trend we will call the "applied approach". This approach developed both in the area of translation studies, where translation theorists saw advantages in analysing the features of texts in order to

provide guidelines for the translation of specific text types, and, more recently, in the area of NLP and AI. In the latter, some of the main aims are to provide a means for analysing text structure and grammar and for producing models of text production and comprehension by computer. This has led to the development of studies in information retrieval and hypertext which combine the two disciplines of computer science and applied linguistics. As is commonly the case with different approaches to the same task, there has been little communication or collaboration between the applied and theoretical approaches with the result that progress has been slow. In the following section we will mention some of the main works in both the theoretical and applied approaches to textual analysis. As it would be impossible to represent all the work carried out in this area here, we will mention only the most influential works along with those that are most relevant to this study.

1.4.2.1 Theoretical Approach

In the area of "Text Linguistics" de Beaugrande and Dressler identified "Seven Standards of Textuality" which every text must have in order to communicate efficiently.[25] Halliday and Hasan studied the use of cohesion in the English language during the late seventies, while van Dijk concentrated on macrostructure and information structure of texts.[26] Biber used a statistical analysis (or "cluster analysis") to examine the notion of text genre and text type across spoken and written language.[27] He maintains that the functional analysis of text types has not been successful and that a text typology should be based on the analysis of co-occurrence restrictions on lexical and syntactic features. He also holds the view, frequently echoed in the analysis of text types for translation purposes, that texts cannot be divided into

---

[25] Dressler, W. & R. de Beaugrande, *Introduction to Text Linguistics*, (London & New York: Longman, 1981).

[26] Halliday, M.A.K. & R. Hasan, *Cohesion in English*, (London and New York: Longman, 1976.
van Dijk, T., *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*, (London and New York: Longman, 1977).

[27] Biber, D., *Variation across Speech and Writing*, (Cambridge & New York: Cambridge University Press, 1988).

sharply distinct "types". Instead there is a continuous range of variation in linguistic form and use. Differentiating between "genre" and "type" Biber states that

> In a fully developed typology of texts, genres and text types will be distinguished, and the relations among and between them will be identified and explained.[28]

Here "genre" refers to categorisations assigned on the basis of external criteria and "type" refers to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories.

Efforts to provide descriptions of text grammars have included TSPs (Texts for Specific Purposes) as well as TGPs (Texts for General Purposes). To date, descriptions of the characteristics of texts in restricted domains has largely come from the German school. As previously stated, Hoffmann is one of the pioneers in this field[29]. Like Biber, Hoffmann believes in using a statistics-oriented approach for the classification of text types. His "cumulative analysis" approach, which involves examining everything from macro-structure to syntax and lexicon, to grammatical category and morphemes, produces two matrices: a structural matrix which analyses macrostructure, coherence, syntax, lexicon and grammatical category, and a functional matrix which examines situation (i.e. social context), intention, communicative situation (e.g. the medium through which communication is established) and the subject domain. Hoffmann believes that, in particular, an analysis of the macrostructure of texts for specific purposes (TSPs) will allow greater insight into the processes of text production and reception. For him, the macrostructure reflects the cognitive division of the text into different concepts which, on the surface, is seen as a linear progression of the parts of a text. However, he concedes that while macrostructure and coherence are essential differentiating characteristics of TSPs, they are not sufficient on their own for classifying text types:

---

[28] Biber, D., op cit, p. 207.
[29] Hoffmann, L., op cit, 1984.

> Sie müssen [...] durch weitere strukturell-linguistische und funktionell-kommunikative Parameter ergänzt werden, so daß sich ein möglichst vollständiges Bild vom jeweiligen Fachtext als strukturelle und funktionale Ganzheit ergibt.[30]

Finally, Hoffmann insists that any classification of sublanguages and their relevant text types should be on the two planes mentioned earlier, i.e. that of the horizontal, where they are classified according to their content and linguistic features, and that of the vertical, where they are classified according to such criteria as situation, channel, receiver and the nature of the sub-domain (e.g. highly abstract or concrete).

Baumann supports Hoffmann's theory that macrostructure is a crucial consideration in the classification of TSPs.[31] His opinion is that macrostructure can be approached from three different viewpoints in text linguistic theory: (i) semantic, (ii) pragmatic and (iii) functional-communicative. Some of the factors to be considered in these approaches include the intention of the text author, the logical content of the subject, the type of communication, the social position of communication partners and the common knowledge and presuppositions of communication partners. However, Baumann points out that textual analysis cannot be carried out on the basis of the pragmatic approach alone because the macrostructure of a text cannot be determined solely on the basis of external factors. It is also a function of the semantic structure of the text. Like Hoffmann, Baumann considers the classification of text segments into some sort of hierarchical structure. He sees the classification of these sections of texts as a basis for determining what text type a particular text belongs to. This is based on the hypothesis that devices used to signal new sections of text have text type specific characteristics.

---

[30] Hoffmann, L., "Makrostruktur und Kohärenz als Fachtextsortenmerkmale", in *Wissenschaftliche Zeitschrift der Karl-Marx Universität Leipzig*, (Gesellsch.-wiss. Reihe, no 37, 1988), pp.552-565 (p.563).

[31] Baumann, K., "Die Makrostruktur von Fachtexten", in *Special Language - Fachsprache*, (vol 1-2, 1987, pp. 2-18).

Gnutzmann and Oldenburg are also proponents of the empirical approach to the analysis of LSP texts.[32] They identify three long-term objectives of TSP analysis which are:

(i) to develop a comprehensive and practical framework for the analysis of LSP texts

(ii) to develop a scientifically-based and comprehensive typology of LSP texts

(iii) to make a significant contribution to the description of special languages which are still rather incomplete because of the long-standing neglect at the text level

Because there is no generally agreed approach to the analysis of LSP texts, Gnutzmann and Oldenburg develop their own framework which draws on several existing models, including those of the aforementioned Hoffmann and Baumann. This model aims at

> facilitating a comprehensive analysis and description of text structures including the crucial relationship between the linguistic form and the communicative function(s) of the LSP texts.[33]

Their model consists of three fundamental steps:

Step 1: select texts belonging to the same horizontal and vertical level in the classification of LSP texts as well as to the same text type.

Step 2: Phase 1 - analysis of text macrostructure

Phase 2 - analysis of text segments

---

[32]Gnutzmann, C., & H. Oldenburg, "Contrastive Text Linguistics in LSP-Research: Theoretical Considerations and some Preliminary Findings", in Schröder (ed.), *Subject-oriented Texts - Languages for Special Purposes and Text Theory*, (Berlin: de Gruyter,1991), pp. 103-136.

[33]Gnutzmann, C. & H. Oldenburg, op cit, p.111.

Phase 3 - microanalysis of text segments

These three phases are interdependent.

Step 3: Evaluation of research results.

Schröder reviews the trends in text-linguistic research of LSP texts (or *Fachtextlinguistik* as he calls it) since the 1970s. He comments on the lack of a systematic development of a theory of LSP texts (mainly due to a lack of a systematic development in text theory in general) and calls for a broadening of the traditional text-linguistic framework towards a *semiotic textology* which, he believes, is necessary because neither text grammar, nor text linguistics, nor the theory of text interpretation can account for all aspects of textuality[34] Thus, a semiotic approach to TSPs would take the subject field, the communicative situation, the scientific paradigms and the cultural input into science and technology into account, as well as the linguistic features of the text:

> in further research we have to take into more serious consideration the level of the content and the communication situation of special texts. [...] Furthermore, we have to analyse the whole complexity of (LSP) text processing, that means the production and reception of special texts within one culture and between various cultures.[35]

He calls for an increase in contrastive analysis research as well as in a consideration of the problems caused by cross-cultural (special) communication.

Other elements which have been more or less ignored in the past, according to Schröder, are the cognitive aspects of text processing and the level of non-linguistic means of textualisation. However, these issues are now being addressed by researchers in artificial intelligence and NLP. Furthermore, the focus to date has been on texts in

[34]Schröder, H., "Linguistic and Text-theoretical Research on Languages for Special Purposes - A Thematic and Bibliographical Guide", in Schröder, H. (ed.), op cit, pp.1-48.
[35]Schröder, H., op cit, p.27.

science and technology (and mainly on explicative, descriptive and instructive text types within those domains) and the sublanguages of social science have received little attention. This trend also seems to be changing with increasing interest in the sublanguages of social science. To conclude, Schröder is convinced that only a semiotic textology can become the general methodological framework for the required research because semiotic textology represents an *integrative* theoretical perspective.

In their consideration of "English Special Languages", Sager et al describe a model of traditional text forms which are defined primarily by the following:

- subtype of one or more major message types
- technique of communication
- status and knowledge relationship among participants
- other pragmatic or linguistic features[36]

They categorise the major text types and define the traditional forms in communicative situations. One of the forms considered is "instructions" which, they maintain, can contain directive, informative and even evaluative information. Rather than consider instructions to be imperatives for action, as is normally the case, Sager et al consider them to be indications of the ways in which actions are to be carried out if the recipient decides to undertake the actions. They believe that the form varies according to the level of special knowledge of the recipient. This view is also held by Hatim and Mason (see further on) and will be further validated by the analysis of our special text types in the next section.

[36]Sager, J. C. et al, op cit, p.148.

1.4.2.2 The "Applied" Approach: Textual Analysis and Translation Studies

As far back as the earliest Bible translations, debates were going on as to whether translation should be a word-based or meaning-based process, the latter method placing emphasis on the text as opposed to the word as the translation unit. Despite these early debates, it was only in the mid-1940s that the first attempts were made to classify texts for translation purposes.[37] Following the growing interest in the text as a linguistic unit in general, there followed several attempts to classify texts for translation in the 1950s and '60s.

Wilss outlines the difference between a text linguistic theory in general and a text-linguistic-based translation theory. The aim of the former is to expound the grammatical macro- and micro-structures of texts, to analyse coherence and other text linguistic features while the aim of the latter is to recommend methods for the translation of specific text types based on an empirical investigation and cross-linguistic comparison of text-type specific features. Wilss maintains that if any progress is to be made in this discipline, both the theoretical and empirical approaches must merge and become one discipline.

Wilss endorses the assumption of scholars such as Coseriu, Nida and Reiss that a text can express three basic functions. These are expressive, informative and imperative. However, he questions whether it is possible to adequately describe every text type according to only one of these functions when it has been demonstrated on several occasions by different researchers that text types are multi-functional.

The notion of multi-functionality in the context of texts for translation is also addressed by Stolze.[38] Stolze agrees with Wilss that it is impossible to attribute one function to a

---

[37] See Wilss, W., *Übersetzungswissenschaft: Probleme und Methoden*, (Stuttgart: Ernst Klett Verlag, 1977).

[38] Stolze, R., *Grundlagen der Textübersetzen*, (Heidelberg: Julius Groos Verlag, 1982).

particular text type. Further, the property of "multi-functionality" can be attributed to individual texts as well as particular text types:

> nicht nur Texttypen sind multiperspektivisch, sondern jeder Einzeltext ist multiperspektivisch angelegt.[39]

Stolze also claims that it would be much more useful for the translator to examine the author's use of language with the aim of establishing the author's intention, rather than list the characteristics of particular text types and link them to function (as Reiss does[40]). In other words, text function should be seen as equivalent to author's intention:

> Die so verstandene Textfunktion erscheint dann vorläufig als Ausdruck der Intention des Textsenders.[41]

Stolze then takes this idea one step further by maintaining that text function has to do with the receiver as well as the producer. Following an analysis of three well-established text types, Stolze comes to the very interesting conclusion that within the so-called text type of "operating instructions" (Bedienungsanleitungen) there are several different text types:

> Offenbar sind nicht alle Bedienungsanleitungen nach demselben Textkompositionsmuster verfertigt, was auf eine Inhomogenität innerhalb der doch weithin akzeptierten Textsorte Bedienungsanleitung schließen läßt.[42]

This discovery has serious implications for any efforts to produce a typology of texts, be it for human translators, for machine translation or for theoretical text linguistics in general. What it implies is that within the traditionally accepted text types, such as "instructive" or "informative" etc., there are further classifications to be made. The features which differentiate these "sub-(text)types" are pragmatic (e.g. previous world

---

[39] Stolze, R., op cit, p. 113.

[40] Reiss, K., *Texttyp und Übersetzungsmethode. Der operative Text*, (Scriptor, Kronberg/Ts, 1976).

[41] Stolze, R., op cit, p. 116.

[42] Stolze, R., op cit, p. 219.

knowledge, author's assumptions and intentions, environment in which the text occurs, existence of standardised formats etc...) as well as linguistic.

Kussmaul analyses operating instructions in English and German from the viewpoint of Speech Act Theory and, in particular, illocutionary acts.[43] His theory is that there is a formal correspondence between text type and "illocutionary indicators". He concludes from his analysis that there are not as many similarities between text types as had previously been accepted and calls for an alteration of the notion of "text type" to take account of this fact.

Hatim and Mason also examine text type from the translator's point of view. They too address the notion of multi-functionality stating that it is the rule rather than the exception:

> multifunctionality is the rule rather than the exception, and any useful typology of texts will have to be able to accommodate such diversity.[44]

They identify three general types of text: exposition, argumentation and instruction and, like Stolze, comment on the fact that within a text type there may be sub-types. For example, they identify two sub-types under the heading of instructional text types. The first type is instruction with option (e.g. advertising, consumer advice etc...) and the second is instruction without option (e.g. contracts, treaties etc...).

Similarly, Salager-Meyer et al, in an examination of texts on "Medical English", conclude that there exists a *continuum* of

---

[43]Kussmaul, P., "Instruktionen in deutschen und englishchen Bedienungsanleitungen", in Arntz, R., G. Thome & W. Wilss (eds.), *Übersetzungswissenschaft - Ergebnisse und Perspektiven: Festschrift für Wolfram Wilss zum 65. Geburtstag*, (Tübingen: G. Narr, 1990), pp.369-379.

[44]Hatim, B. & I. Mason, *Discourse and the Translator*, (London & New York:Longman, 1990), p.138.

> text types which are best seen in a circle with a common core: The core expands in all directions with various realizations of text types at different levels of specialization and abstraction.[45]

1.4.2.3 Text Linguistic Analysis and NLP

Given the increase in interest in textual analysis of specialised texts for translation purposes, it would be natural to expect a corresponding interest in text in the relatively new discipline of NLP and, in particular, MT. However, as we have already pointed out, research in this area has been scarcer than in any of the aforementioned disciplines. Nevertheless, recent developments in artificial intelligence have awakened interest in the text and it is now believed that analysis of text can reveal more about cognitive models of text production and reception. Indeed, de Beaugrande comments that not until more is found out about cognitive models of text production and reception will the AI approach to MT prove beneficial.[46]

Hauenschild is among those who considers the analysis of the text as a necessity if FAHQT is to be achieved. According to her, new developments in AI mean increased possibilities for the processing of whole texts. She proposes an interlingual-transfer approach to MT so that the textual invariants may be encoded in an interlingua and, thus, remain unchanged in the translation process:

> Such an interlingual component might be conceived of as an augmented form of a semantic network with an explicit indication of referential relations and of the hierarchical as well as systematic interrelations between the themes and sub-themes of a text.[47]

---

[45] Salager-Meyer, F. et al, "Communicative Function and Grammatical Variations in Medical English Scholarly Papers", in Laurén, C. & M. Nordmann (eds.), *Special Language: From Humans thinking to Thinking Machines*, (Clevedon & Philadelphia: Multilingual Matters, 1987), pp.151-160 (p. 157).

[46] de Beaugrande, R., "Text Linguistics and New Applications", in *Annual Review of Applied Linguistics*, vol. 11, pp. 17-41, 1990.

[47] Hauenschild, C., "Discourse Structure - Some Implications for Machine Translation", in Maxwell, D. et al, 1988, pp.145-155 (p. 148).

She also suggests that the transfer component be divided into three - syntactic, semantic and pragmatic - where each part accounts for a different aspect of the structure of a text.

Sager sharply criticises the claim made by some MT systems that they are suitable for the translation of particular text types such as instructions or technical manuals when no efforts have been made to reveal *why* or *how* they are suitable.[48] He praises systems such as METEO, CULT and TITUS which are based on not only a particular text type but also a concrete communicative situation, and well-defined and controlled production circumstances (such as weather forecasts in the case of METEO).

However, while it can be said that TAUM METEO developed a highly successful sublanguage-based MT system, it must also be pointed out that the processing unit in this system is the sentence and not the text. The text was considered at one stage of development in TAUM, and it was admitted that there were distinct advantages in using textual information in the translation process. Nevertheless, the text as a unit of translation was abandoned for "reasons of economy".[49] The text was also abandoned in favour of the sentence in TAUM AVIATION because, as the system designers put it, "we still know too little about discourse analysis techniques to use them efficiently in large scale systems".[50] Paradoxically, many of the translation problems in TAUM resulted from weaknesses in sentential processing where ambiguities mainly caused by conjunctions and nominal compounds could not be resolved.

Thus, even though the sentence has been given precedence over the text in MT research, it would seem that researchers are becoming more aware of the benefits of using the text as a unit of translation and of the limitations imposed by the sentence.

---

[48]Sager, J.C., "Machine Translation and a Typology of Texts", in Laurén, C. & M. Nordmann (eds.), op cit, pp.397-410.

[49]Lehrberger, J, in Kittredge, R. & Lehrberger, J. (eds.), op cit, p.97.

[50]Isabelle, P & L. Bourbeau,"TAUM-AVIATION: Its Technical Features and Some Experimental Results", in *Computational Linguistics*, vol. 11, 1985, pp. 18-27.

For example, Kittredge and Lehrberger believe that the future of machine translation is narrowly linked to developments in textual analysis as well as sublanguage analysis:

> The commercial success of machine translation in the foreseeable future likely depends on the possibility of writing sublanguage grammars for texts in particular fields.[51]

Similarly, Weber argues that LSP (and sublanguage) research must primarily be concerned with the classification and analysis of special-purpose text types because it has been demonstrated that an analysis of syntax and semantics alone is entirely inadequate as an attempt to describe special languages. He states clearly his belief that positive side effects are achieved by specialising MT systems in a particular domain and text type. However, how far textual characteristics are taken into consideration depends on the quality of translation desired.[52]

Lehrberger mentions the existence of text norms for writing recipe books, articles in scientific journals etc, which ultimately means that there are distinct regularities in these text types, thereby making it easier to write text grammars:

> The existence of norms for texts in certain fields, the reduction in polysemy resulting from semantic restrictions, the limited vocabulary, and the syntactic restrictions generally encountered all combine to make automatic translation practicable for sublanguages.[53]

He also considers the relation between text grammar and text purpose, mentioning that an important factor in determining the use of language in a text is the purpose for which the text is intended.[54] Lehrberger observed considerable differences between texts from the same sublanguage domain (i.e. weather forecasting) which had different purposes. One text type was labelled "weather bulletin" and its function was to report the latest forecast as concisely as possible. The other text type was called "weather

---

[51] Kittredge, R. & J. Lehrberger, op cit, p. 3.

[52] Weber, H., *Converging Approaches in Machine Translation: Domain Knowledge and Discourse Knowledge*, (University of Duisberg Linguistic Agency, Series B, Paper no. 164).

[53] Lehrberger, J., op cit, p. 99.

[54] Lehrberger, J., in Grishman & Kittredge (eds.), op cit, p.28.

synopsis". Its purpose was to give a general summary of weather conditions with less emphasis on brevity. Moreover, within the text type "aviation hydraulics manual" there were different sections which were characterised by different linguistic characteristics. On closer inspection, it was noticed that the purpose of one section was to instruct the reader to carry out a procedure (this was characterised by the use of imperatives and short sentences) and the purpose of the other section was to describe an object in the aircraft and/or its function. There was a distinct lack of the imperative in these sections and the sentences were longer. Lehrberger points out that although text purpose generally affects text structure in fairly predictable ways, the relation between purpose and structure is a complex one. It is not simply a matter of a given purpose determining a specific structure. Other criteria must be considered, such as guidelines for the producers of texts and the degree of expertise of the text receiver, to name but two.

The reasons progress has been slow in the area of textual and text type analysis for MT to date are, in our opinion, manifold. Firstly, the process of machine translation itself has proved more difficult than first imagined. For this reason, researchers have concentrated their efforts on finding the most efficient *method* for automatic translation rather than the most suitable sublanguages and text types. Secondly, textual analysis is itself a relatively new discipline and efforts are still being made to develop a unified text theory for both general purpose and special purpose texts. Finally, it is only recently that MT researchers have begun to acknowledge the benefits of considering the text as a unit for machine translation.

1.4.3 Conclusion and Recommendations

We have seen from our brief look at the area of text linguistics and sublanguage/machine translation that, to date, the text has been given little consideration by developers of MT systems. On the other hand, analysis of text has

been going on for many years both in the area of theoretical applied linguistics and in translation studies and NLP. The results of textual analysis in these domains has not gone unnoticed in the domain of MT, and many MT researchers now acknowledge the importance of considering textual characteristics in MT systems. They are in a position where they can draw on the experience and conclusions of previous text linguistic researchers and incorporate the latter's recommendations into the design of future MT systems.

The current trend in MT is to restrict the system to one sublanguage while at the same time trying to ensure extensibility so that the system might be able to deal with other sublanguages in time. However, on the basis of the research completed to date, we know that even if an MT system is restricted to one sublanguage and/or one specific text type, there is no guarantee that each text will have identical linguistic features. The reasons for this are manifold: for example, as Biber points out, there is a lack of *distinct* text types. Rather, text types should be seen as forming part of a continuum with an identical core. This fact applies both to TGPS and TSPs. Also, within one text type there may be several sub-types as Hatim and Mason have demonstrated with the "instructive" text type which may be classified as "instructive with option" and "instructive without option". These sub-types can be identified by the fact that they display different macro- and micro-linguistic characteristics. Finally, while it is generally accepted that there is a correlation between communicative function and linguistic features, it is also acknowledged that most text types, and indeed most texts, are multi-functional with one function dominating in one text segment and another function dominating in a different text segment.

We conclude from this that the sublanguage/text-type approach to MT is not unproblematic. Nevertheless, it does provide for more optimism when compared with the task of translating non-restricted texts automatically because sublanguage grammars are generally more restricted than the grammar of standard language.

Furthermore, processing at the level of the text may eliminate ambiguities which are impossible to eliminate at the sentence level. Moreover, even if there are many text types and sub-types with several communicative functions, the task of identifying them and classifying them is not as huge as for general language texts. To be successful in this approach developers of MT systems should heed the recommendations of previous text researchers. This implies the following:

(i) a statistics-oriented approach should be taken

(ii) co-occurrence restrictions should be analysed, not just frequency of occurrence

(iii) extra-linguistic features should be taken into consideration as well as linguistic features. This requires examining author's intention, communicative situation, channel, receiver status, common knowledge and presuppositions, nature of sub-domain etc.

And finally,

(iv) the textual analysis should be performed at both the micro- and macro-level in order to make use of the important information contained in the macro-structure of the text, something which is normally ignored in sublanguage/MT analysis.

We will now move on to an analysis of text types within a particular sublanguage with a view to establishing how useful text type analysis might be for MT (Chapter 3). Before this, it is necessary to discuss the compilation of corpora for sublanguage research (Chapter 2). The compilation of the sublanguage corpus is a phase which is necessary for every sublanguage project but, not unlike the issue of text type, it has

received little attention in sublanguage literature so far. Following the recommendations made by text linguistic researchers, the extra-linguistic features of the text types in our corpora will be described before turning to a more detailed micro-linguistic analysis in Chapter 3.

CHAPTER 2:

THE SUBLANGUAGE CORPUS

2.0 Introduction

In the area of computational linguistics, there has been a recent upsurge in the computational analysis of large bodies of text. According to Walker, this type of research is "essential for the development of adequate models of linguistic structure and for insights into the nature of language use".[55] To date, corpus linguistics has concentrated mainly on general language corpora. For example, the compilation of the SEU (Survey of (Educated) English Usage) corpus began in 1959, following an observation that there was a lack of a proper basis for the writing of a grammar of English. This corpus used both spoken and written material. Compilation of the London-Lund Corpus, a sister project of the SEU, commenced in 1975 at the University of Lund. Its aim was to make available, in machine-readable format, the spoken material used in the SEU project. The Brown Corpus of American English was the first computer-based corpus to be compiled (1963-64). It consists of one million words. In order to provide a British English counterpart to the American Brown corpus, the Lancaster/Oslo/Bergen (LOB) corpus was compiled. It also consists of approximately one million words. The COBUILD corpus was developed approximately ten years ago at the University of Birmingham with collaboration from Collins and the English Language Research Group at the university. It was compiled for the purposes of lexicography and consists of 20 million words. Other corpora exist but they are too numerous to list individually here.

Although much work is going on in the area of corpus compilation of general language texts, there has been little or no discussion of sublanguage corpora. Admittedly, Kittredge mentions the notion of a "representative" corpus which is used for measuring closure. Nevertheless, no criteria are offered by him for the compilation of a

[55] Walker, D., "Collecting Texts, Tagging Texts, and Putting Texts in Context", in *Working Notes for AAAI Spring Symposium on Text-based Intelligence Systems*, (Stanford University, March 1990), pp. 39-41.

sublanguage corpus. It is necessary that this area be addressed because any sublanguage analysis, be it for MT purposes, for lexicographical purposes or simply for teaching purposes, is based on a corpus of sublanguage texts.

2.1 Uses of Sublanguage Corpora

Sublanguage corpora can be used for several purposes. For example,

(i) Teaching Language for Special Purposes:

A sublanguage corpus provides invaluable information about the lexicon, grammatical structures and co-occurrence restrictions in a specific domain which can then be used as course material for non-native speakers of the sublanguage.

(ii) Lexicography:

Sublanguage corpora can be used to compile specialised dictionaries and glossaries, providing information not only on the terms which occur in a specialised subject-domain but also on the contexts in which they occur.

(iii) Controlled Language:

In the previous chapter it was mentioned that writers of specialised texts often follow specific guidelines in order to produce unambiguous, well-structured texts. As ambiguity is often a problem, even in sublanguage texts, analysis of corpora of texts from specific sublanguages would reveal recurring ambiguous structures. The writers of such texts could then be trained to eliminate ambiguities so as to ease comprehension and to facilitate the translation process.

(iv) Documenting of Language Change and Subject-field Development:

Sublanguage corpora can be used to chart the diachronic development of the language used in a subject-specific area. This would provide etymological information on terms and concepts and could even be used in sociological studies for monitoring the development of a particular science or technology.

(v) Natural Language Processing:

Machine-readable sublanguage corpora could be used for CALL applications ("Computer Aided Language Learning") or even for CAL applications ("Computer Aided Learning"). On-line specialised information could also benefit translators, terminologists, technical writers, editors and so on.

Bi-lingual sublanguage corpora can be used for comparison of text types, syntactic structures and terminology for the development of sublanguage-based machine translation systems. In this context, Sebba mentions three applications of sublanguage corpora:

(i) monolingual corpora provide a database of lexical information and a basis for automatic parsing for the source language (SL). They are also used to develop tools for MAT (Machine-Aided Translation), e.g. dictionary production.

(ii) Bilingual corpora can be used for translation from SL text to TL text.

(iii) Corpora can be used for testing experimental systems.[56]

[56]Sebba, M., "The Adequacy of Corpora in MT", in *Applied Computer Translation*, (Vol. 1, Issue 1, 1990), pp. 15-27

While Sebba concentrates on the actual use of corpora in the development of MT systems and mentions a specialised sublanguage corpus (service manuals) which is used as a database for an MT system, no indication is given about the criteria used for compiling this corpus or any other sublanguage corpus.

2.2 "Representativeness"

The function of any corpus is to be representative. For example, it can be representative of linguistic variation within a specific subject domain across languages, or of the language used in a specific domain in one language, or of a particular text type, and so on. However, no matter how large a corpus is, it cannot be ensured that it is fully representative of the phenomenon it seeks to describe. Sebba maintains that, given a large enough random sample, there is a statistically good chance that it is representative but there is no guarantee of this. Bungarten refers to a large corpus of randomly selected texts as being exemplary, and believes that an exemplary corpus is more easily attainable than a representative corpus.[57]

The likelihood of a sublanguage corpus being representative is much higher than a standard language corpus because the universe which the sublanguage describes is finite. Kittredge rightly points out that whether or not a sublanguage corpus is representative depends on the way in which the boundaries of the sublanguage are defined. For him, a representative corpus is related more to the description of the sublanguage grammar than to the lexicon. What is required of a sublanguage corpus is

[57]Bungarten, T., "Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft", in *Empirische Textwissenschaft*, Bergenholtz, H. & B. Schaeder (eds.), (Königstein/Ts. Scriptor, 1979).

> A large enough view of the sublanguage to set up all the necessary categories for its description, and all the admissible sequences of categories which are to be admitted as sentences.[58]

Thus, a sublanguage corpus would allow all possible relations between nouns, verbs and their modifiers to be discerned. Once these relations have been established, any new lexical items should fit into the already existing categories.

Atkins, Clear and Ostler present the notion of a "balanced" corpus and present general guidelines for the compilation of such a corpus.[59] To attain a balanced corpus, the corpus builder must first attempt to build a representative corpus which is then analysed for strengths and weaknesses. In the light of this analysis, the corpus is enhanced by the addition or subtraction of material and the cycle is repeated continually. This process is one of successive approximation which improves the balance of the corpus to suit the needs of the users. In an ideal situation, this would obviously be a very beneficial process. However, given the time constraints of modern R&D programmes, trade-offs between fully representative or balanced corpora and exemplary corpora are often necessary.

2.3 Sublanguage Corpus Development

Any MT system will require contrastive analysis during the research stages and this, in turn, will require the compilation of bilingual corpora. Although the composition of corpora very much depends on the individual aim(s) of the research project, there are surely some benefits to be gained from establishing a set of *general* guidelines for building corpora. Some of the issues to be considered here are:

---

[58] Kittredge, R., op cit, 1982, p.130.

[59] Atkins, S., J. Clear & N. Ostler, "Corpus Design Criteria", in *Literary and Linguistic Computing*, (Vol. 7, no. 2, 1992), pp.1-16.

- What expertise is required for the compilation of sublanguage corpora?
- Corpus size
- Text selection
- Authorship
- Format
- Copyright

Each of these issues is considered individually below. However, it must be reiterated that criteria for the compilation of corpora very much depends on the aim of the individual research project. Therefore, the guidelines that follow are intended to be general.

What expertise is required for the compilation of sublanguage corpora?

Firstly, it is necessary that the compilation of a sublanguage corpus involve an expert in the sublanguage working in conjunction with sublanguage/MT researchers. There are a number of reasons for this:

(i) It is possible for an educated lay-person to recognise and understand highly specialised domains, but only an expert can estimate the closure level of the domain.

(ii) An expert can indicate the possible range of text types that occur in a sublanguage.

(iii) An expert is usually familiar with the development of a subject area on a global scale. This is an important factor when developing a bilingual corpus because the terminology of a specialised domain may be more developed in one language than in another.

(iv) An expert can indicate whether a specialised field is likely to develop rapidly or not. The advantages of compiling a corpus of texts from a rapidly developing field of science or technology are limited as the terminology related to that particular domain will still be evolving. Along with an expert in the subject domain, computational linguistic expertise is also required as computational experts will know what they aim to achieve by compiling the corpus and what the MT system is capable of.

Corpus Size

The size of corpora compiled to date varies from a few thousand words to a few million, and even a billion, words. If the aim of the project is to monitor linguistic variation in a particular language, the corpus must be very large, i.e. a million or more words. A collection of the biggest English language corpora compiled to date is now available in CD-ROM format from the International Computer Archive of Modern English (ICAME) which resides at the Norwegian Computing Centre for the Humanities (NCCH) in Bergen. This CD-ROM includes the Brown corpus, the LOB corpus, the Kolhapur corpus, the London-Lund corpus and the Helsinki corpus and is available for a nominal sum to anybody wishing to carry out linguistic research on some aspect of the English language. On the other hand, if the aim is to develop a MT system for a particular sublanguage, as in the case of TAUM METEO, the corpus must be large enough so as to provide examples of all possible syntactic constructions and semantic categories. The size of a corpus with this purpose depends on the size of the sublanguage. Moreover, if the aim is to carry out research on a sample of sublanguage texts or text types then the corpus will be tailored accordingly to suit the aims of the research.

Text Selection

As our analysis in Chapter 3 will reveal, a sublanguage may consist of many text types with different linguistic and extra-linguistic features. If it is the aim of a corpus to describe a sublanguage, then, obviously, all text types must be included. As Biber points out, this will necessitate considerable pre-analysis to

> identify (1) the parameters of situational variation within this domain; (2) the different processing constraints within this domain; (3) the different communicative tasks in this domain; and (4) the different relationships among communicative participants.[60]

Thus, in any corpus which seeks to be representative of a sublanguage, all possible communicative situations, and their corresponding text types, should be included. When compiling both bilingual and monolingual corpora, efforts must be made to ensure that *parallel* sublanguage texts are included and not *translated* texts.

Authorship

The author(s) of the texts should be easily identifiable and the source of all texts should be known and well documented so that they can be traced if necessary. When creating a bilingual corpus it is necessary that the text producers are native speakers of each language to ensure correct use of terminology and grammatical structures.

Format

The lack of machine-readable material in the past meant long hours sitting in front of a computer keying in text before any linguistic analysis could take place! Moreover, due

---

[60] Biber, D., "Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses", in *Linguistics*, (Vol. 23, 1985), pp.337-360 (p.341).

to an absence in standardised information formats, the input material could be used only by those working on that particular project. Few, if any, other research groups could avail of the machine-readable information. With the boom in information technology much more material is now available in machine-readable format. The problem of the absence of a standard format is also being addressed with the development of SGML - Standard Generalized Markup Language- (ISO 8879) which is a modelling language used to define the structure of a document. Encoding documents with SGML tags facilitates the transfer of those documents from computer to computer.

> The techniques provided in the standard allow computers with different character sets and coding schemes to communicate in an internationally agreed way. Once this is possible, the barriers between computer systems can be removed and information can, at last, flow freely around the world.[61]

Documents can still be typed in traditional word processing packages (for example, Word Perfect) and then these can be tagged using SGML. This is an obvious advantage for the collation of texts, not only for MT, but for all linguistic research projects and efforts should be made to ensure that the standard is used.

Copyright

One of the practical problems which has to be overcome when compiling any corpus is that of copyright. A project may have reached the stages of production when the researchers discover that they cannot publish or use any of their results because of copyright problems. Payment for copyright and for the administrative work involved in identifying copyright holders can amount to substantial amounts of money, something which research groups often cannot afford. The issue of copyright should be sorted out at the beginning of any project so as to avoid problems at a later stage.

---

[61] Bryan, M., *SGML: An Author's Guide to the Standard Generalized Markup Language*, (Wokingham & Massachussetts: Addison-Wesley, 1989), p. viii.

2.4 Description of the Corpus and Aims of the Analysis

The sublanguage we have chosen for our analysis is that of ***Sewing***.[62] The domain of sewing is relatively closed in that there is a finite number of garments that can be made (although, admittedly, new names can be given to the same garment over a period of time) and a finite number of actions which must be carried out in order to construct a garment. In this respect, it is a good candidate for machine translation. Moreover, sewing is something that is practised in every country and this is reflected by the fact that instructions on sewing patterns are translated into several different languages. Thus, if a MT system were to be developed for the translation of sewing instructions, there would at least be a market for it.

The sublanguage of sewing exhibits all the characteristics of a sublanguage as outlined in Chapter 1. For example, it has a limited subject matter. The language used is a subset of the natural language of English. The texts in our corpora exhibit a certain degree of lexical closure.[63] Grammatical closure, on the other hand, is more difficult to estimate as this would necessitate analysis of a closed corpus for all possible syntactic constructions, something which goes beyond both the aims and the scope of this study. The sublanguage exhibits lexical, syntactic and semantic restrictions. Its complete lexicon would exclude a large portion of the lexicon of English; it does not have the creative potential of, for example, poetic texts; there is a high degree of recurrence of

[62]The choice of sublanguage resulted from the work carried out under the Eurotra framework where the initial task was to analyse texts from the domain of Satellite Telecommunications. It soon became apparent that these texts, although taken from a restricted sublanguage, contained several ambiguities which would cause problems during machine translation. We then set out to find another set of sublanguage texts which could be compared with the Telecommunications texts in order to demonstrate that the latter were not restricted enough for MT. Thus, we chose texts from the sublanguage of sewing which appeared to be (and later were proven to be) much more restricted than the initial corpus of Telecommunications texts.

[63]This phenomenon will be demonstrated in section 3.1.2 in Chapter 3.

certain syntactic constructions, e.g. one-word imperative sentences such as *Press.*, and an absence of questions and exclamations; and, finally, semantic restrictions are exhibited by words such as *tack* which always refers to the process of sewing short, straight stitches and not to the metal object used to fasten notices to notice boards etc., *stitch* which excludes the medical meaning of the term, and *dart* which always refers to a fold made in material and not the object used in a game. Similar to Lehrberger's example of "deviant" rules of grammar in the sublanguage of aviation hydraulics manuals, the sublanguage of sewing has a high denstiy of occurrence of ellipsis, e.g., *Shaping lapel over hand, padstitch lightly front roll line to seamline. Trim interfacing close to padstitching.* It also has a high frequency of certain constructions. For example, there is a very high occurrence of the imperative verb, and a high recurrence of phrases such as *With wrong sides together, do X, With right sides together, do Y*, or *matching X and Y,* to name but a few. Special symbols are used, most of which signify measurement, e.g. *6mm*, *1/4 "*. And, finally, the texts chosen have distinct, repetitive structures which are described in detail below.

Three different "types" of text were chosen for the analysis. The reason for this lies in the aims of the analysis which are twofold:

(i) *To illustrate that within one sublanguage there is frequently more than one text type and that each text type differs linguistically from the next according to its function and other extra-linguistic criteria, and to demonstrate that a text type/text function analysis would be beneficial for MT.*

(ii) *To illustrate that although we can talk of "text types" and "text function" there is no one rigidly defined "instructive" text type. Instructive text types can differ linguistically from each other due to the influence of extra-linguistic criteria such as communicative situation, knowledge level of recipients etc.. It would therefore be wise to include an analysis of extra-linguistic criteria in any text type analysis.*

As was mentioned in Chapter 1, it has been recommended by researchers in the area of text type analysis that macrostructure and extra-linguistic characteristics be considered

in any textual analysis. In keeping with this recommendation, there follows a general description of the texts which make up our corpora as well as a description of their macrostructures.[64]

2.4.1 Corpus 1: Sewing Patterns

The text type in corpus one is a manual of sewing instructions which accompanies pattern pieces. This text type is normally bought in specialised shops which deal with fabrics and needlework. Its primary function is to instruct the expert user in the construction of a garment. The text is written by an expert. Therefore, the relationship between the communicative participants can be described as expert to expert. The written text is accompanied by diagrams which play an important part in the communication process. The text begins with the name of a section of the garment as a title, e.g. *Right Front*, and then proceeds in steps, e.g. *Step 1 - STAY-STITCH upper edge of back pieces*, and continues until that section is completed.[65]

2.4.2 Corpus 2: Pattern Envelope

The second text type is the text contained on the **back** of the envelope which contains the sewing instructions (i.e. the text type in corpus 1) and the pattern pieces. It consists of four distinct sections. The first section is a description of the garment, e.g. *Jacket with funnel neck has princess seaming and long two-piece sleeves.* Section two lists the types of fabric that are suitable for that particular garment. Section three is a list of "notions" or extras such as zip, buttons etc. which are necessary for completing the garment. Finally, section four is a table of body measurements and the corresponding amount of material required to construct the garment. The language in this text type is very telegraphic in nature with few complete sentences and many tables of figures. Like

[64] All three text types are in "written to be read" format.

[65] See Appendix A for a sample of the Sewing Instructions text.

text type one, it is produced by an expert for an expert and its main function is to instruct the user to buy the correct fabric, the correct amount of fabric and the correct size of buttons and zip etc. to finish the garment. Its secondary function is to provide a description of the garment. We consider this to be "secondary" because the main description of the garment is provided on the front of the envelope in the form of a photograph or drawing of the garment.[66]

2.4.3 Corpus 3: Textbook

The text for the third corpus is taken from a section on sewing in a "Home Economics" textbook which is used in the classroom for children aged between 15 and 18. The texts are taken from the section dealing specifically with types of stitches used in the sewing process. The text is written by an expert for the non-expert and its main function is didactic. However, the text also seeks to teach by *instructing* the text recipient to carry out certain functions. It can, therefore, also be called an "instructive" text type.

Each section begins with the name of a particular stitch as the title. Next, a description of the uses of that particular stitch is provided. Finally, the recipient is instructed on how to produce this stitch by following a series of steps.[67]

[66] See Appendix C for a sample of the Pattern Envelope text.
[67] See Appendix B for a sample of the Textbook text.

2.5 Criteria for the Compilation of the Corpora:

The criteria which were recommended above for the compilation of sublanguage corpora were followed as closely as possible when compiling our three corpora. Thus, for example, they were compiled by a sublanguage expert who is also a native speaker of English.

The size of the corpora was a matter of some consideration. At first, corpora of approximately 2 500 words were compiled for both the Sewing Instructions and the Textbook. In the case of the Sewing Instructions, this yielded 301 unique words. And in the case of the Textbook, there were 544 unique words. Another (approximate) 2 500 words were then added to each corpus. The increase in the number of unique words in the Sewing Instructions was quite small, yielding only an extra 117 words to reach a total of 418. In the Textbook, there was an increase of 256 new words when the extra 2 500 words were added (yielding 800 unique words in total). We conclude from these figures that increasing the number of words results in a small increase in the number of unique words. We expect that, eventually, once all garment types and sewing processes are included, there would be no increase in the lexicon. However, as our aim is not to describe the whole sublanguage, but to examine some linguistic features of some text types in the sublanguage, we felt that approximately 5 000 words was an adequate sample to complete this task. Other issues involved in deciding what size the corpora would be were time and financial constraints. Financial constraints limited the number of patterns we could buy. Moreover, since none of the corpora were in machine-readable format (SGML is not so popular yet that sewing instructions are marked using it!), it was necessary to key in all texts ourselves. This is a time-consuming task which necessitated limiting the number of words in each corpus. Since the Pattern Envelope corpus is extremely telegraphic in nature and consists of many

tables of figures it would have taken too many pattern envelopes to attain 5000 words of text. To reach approximately 5 000 words of sewing instructions, five patterns were required. Therefore, it was decided to use the "Pattern Envelope" text on the back of each of these five patterns. The total number of words in these five texts amounts to 1335. Overall, the number of words in each corpus provides an adequate example of the linguistic nature of each text type.

For the Sewing Instructions and the Pattern Envelope texts, the authors are not easily identifiable because the patterns are produced under a particular pattern brand name and the individual authors are not identified. Texts from two of the leading pattern-producing companies in the English speaking world were chosen. The authors of the Textbook texts, on the other hand, are easily identifiable as their names are provided on the cover of the textbook.

Copyright did not cause any problems because the texts were used for research purposes only. However, if a machine translation system were to be developed for this sublanguage, it would be wise to investigate this matter before proceeding with the project.

Following the conversion of the texts into machine-readable format, they were indexed and saved into a text indexing and retrieval package called *Wordcruncher*. This made it possible to view words in context, to make lists of occurrences and to view the frequency with which lexical items occurred. Nevertheless, much manual analysis was also required.

2.6 Summary

In this Chapter we have reviewed the situation of sublanguage corpora in general, commenting on the uses of sublanguage corpora and on the fact that no guidelines have yet been offered for the compilation of sublanguage corpora for machine translation purposes. This is likely due to the fact that criteria for the compilation of corpora depend on the aims of the individual research projects. With the advent of an age where machine-readable material can be marked into a generalised format and exchanged easily by computer, it is necessary that at least general criteria for the compilation of corpora are agreed. With this in mind, we have suggested some criteria which might aid in the compilation of sublanguage corpora in the future, while all the time emphasising that criteria ultimately depend on each individual project.

Finally, in keeping with the recommendations of text (linguistic) researchers, we have provided a description of the extra-linguistic features of each of the three text types which make up our corpora. We will now move on to the analysis of these text types.

CHAPTER 3 :

ANALYSIS

3.0 Introduction

This chapter presents an analysis of the three corpora and the results obtained. Since the discussion focuses on the sublanguage approach to MT, the texts are analysed for linguistic phenomena which cause ambiguity and which render automatic translation problematic. Three types of ambiguity - lexical, semantic and syntactic - and two linguistic features under each of these headings are examined (Section 3.3). This is preceded by a general discussion of linguistic phenomena which cause ambiguity and the methods used to solve these problems (Section 3.2). Before this, however, the corpora are compared for lexical closure (Section 3.1 below).

3.1 Lexical Closure

3.1.1 Introduction

As mentioned in the introduction, Zellig Harris (1968) proposed that one of the main properties of sublanguages is that they have *closure*. According to Moskovich closure means that

> If we have a finite set of texts belonging to a certain sublanguage, we can determine its vocabulary and its inventory of grammatical constructions.[68]

In other words, if we have a set of texts from a sublanguage which contain all possible grammatical constructions and co-occurrence restrictions and which demonstrate how many grammatical functions and grammatical categories a particular lexical item can have, then the addition of extra texts will not cause the grammar or the vocabulary to

[68]Moskovich, W., "What is a sublanguage? The notion of sublanguage in modern Soviet linguistics", in Kittredge & Lehrberger (eds.), op cit, pp.191-205 (p. 193).

change (unless there has been a new discovery in that area and a new word has been coined to describe that discovery). The degree of closure in every sublanguage depends on the "closedness" of the semantic domain it seeks to describe. In general, sublanguages are more or less closed while "language as a whole" is an open system.

3.1.2 Degrees of Lexical Closure

While an examination of grammatical closure in sublanguages is beyond the scope of this study, it is possible to compare two sublanguages for degrees of lexical closure. To demonstrate this we compared approximately 5000 words of the sublanguage of sewing instructions with approximately the same number of words of the sublanguage of satellite telecommunications.[69] The text indexing and retrieval system ("Wordcruncher") counted the total number of words in each corpus and the total number of *unique* words. The total number of words in the Telecommunications corpus is 5107 of which 1362 are unique. In other words, those 1362 words are repeated to create text which comprises 5107 total words. The total number of words in the Sewing Instructions corpus, as already mentioned, amounts to 5232, slightly higher than the Telecommunications corpus. Of these 5232 words only 418 (in comparison with 1362 in the Telecommunications corpus) are unique. This implies that the corpus of Sewing Instructions is more restricted lexically than that of Satellite Telecommunications.

It is our opinion that closure is a property not only of the sublanguage itself but also of the different text types which may occur in any sublanguage. To demonstrate our claim that individual text types within sublanguages may also display different degrees of

[69]This analysis was carried out by the members of the Eurotra Ireland team during the research programme on Sublanguage and Text Type which was mentioned earlier.

lexical closure, we compared the three corpora of text types in the sublanguage of sewing, i.e. the Sewing Instructions, the Pattern Envelope and the didactic Textbook. As mentioned above, the corpus of Sewing Instructions amounts to 5232 words, of which 418 are unique. The corpus of texts from the didactic Textbook on sewing has 5274 words in total, of which 800 are unique. Hence, although the two corpora have almost an equal number of words, the Textbook corpus has almost double the number of unique words, implying that the Textbook is less closed lexically than the Sewing Instructions.

The corpus of texts from the Pattern Envelope amounts to 1335 words with 343 unique words. When we compare this corpus with that of Sewing Instructions we conclude that the latter is much more closed lexically as it has only 418 unique words while the total number of words amounts to 5232 (in comparison with 1335 for the Envelope corpus).

By examining our corpora for lexical closure we have demonstrated that different text types in the same sublanguage can display different degrees of closure as is illustrated by the table below:

Table 3.1: Closure

Sublanguage	**Sewing Instructions**	**Textbook**	**Envelope**
Total no. Words	**5232**	**5274**	**1335**
No. Unique Words	**418**	**800**	**343**

We will now move on to a general discussion of the phenomena which cause ambiguity for MT and the methods by which this problem can be resolved.

3.2 Ambiguity and Machine Translation

3.2.1 Introduction

Analysis of literature on machine translation reveals that all systems have to tackle similar linguistic problems which cause ambiguity. These linguistic ambiguities can be divided into three categories: lexical, semantic and syntactic.

Lexical ambiguities can be sub-divided into three types: categorial ambiguities, homographs and polysemes, and transfer or "translational" ambiguities.[70] Categorial ambiguities arise when a word is assigned to more than one grammatical or syntactic category, something which occurs quite often, and the problem is accentuated when several categorially ambiguous words occur in the same sentence, e.g. *Foot heads arms body.*[71] To disambiguate this example, world knowledge along with knowledge of the structures typically used in newspaper headlines is necessary.

The term "homograph" is generally understood to mean lexical items which are spelt the same but which have very different meanings, e.g. ***club***, while "polyseme" denotes a lexical item which exhibits a range of meanings all of which are related in some way to each other, e.g. *bridge* (the dental kind, the structural kind etc.). There has been much debate between linguists about the difference in meaning between the two terms

[70]Hutchins, W.J., & H.L. Somers, *An Introduction to MT*, (London: Academic Press, 1992), p. 99.
[71]Hutchins, W.J. & H.L. Somers, op cit, p.86.

and it is generally understood that a lexical item which is considered to be a polyseme can, with the passage of time, be considered a homograph. The distinction is irrelevant for machine translation because the implications, i.e. that there can be more than one translation of the lexical item in the target language, are the same whether the lexical item is considered to be a polyseme or a homograph. For this reason the term "polyseme" will be used in this analysis to denote any word which has a range of different meanings.

It is sometimes the case that a source language word can be translated into several target language words or expressions depending, of course, on the meaning and the subject domain. This is what is known as a "translational" or transfer ambiguity. Since our analysis is monolingual, transfer ambiguities will not be considered.

Syntactic ambiguity arises when there is more than one way of analysing the underlying structure of a sentence. It is caused by phenomena such as prepositional phrase (PP) attachment, conjunction, ellipsis of articles, embedded clauses and adjective scope.

Finally, semantic ambiguity is caused by synonymy, long nominal compounds and referential items whose antecedents are not easily identified.

3.2.2 Resolution of ambiguity

In machine translation, linguistic ambiguities can be resolved using several different strategies. For example, parsers can be provided with linguistic knowledge about co-occurrence restrictions, i.e. indications about the likelihood of occurrence of certain elements given the occurrence of other elements in the sentence. Given that sublanguages are more restricted in nature than standard language, outlining the co-

occurrence restrictions for a sublanguage domain should be easier than for a standard language domain. Secondly, many sentence ambiguities can be disambiguated if reference is made to the context in which they occur. At the moment, however, this is not an option for machine translation because the information necessary for disambiguation might be found in the previous sentence or in the previous chapter. There are no hard and fast rules about where to locate the piece of knowledge necessary for disambiguation and, for this reason, the use of contextual information for disambiguation is highly restricted. Moreover, it would be impossible to store every possible interpretation of every sentence in every possible context in a machine translation dictionary. The use of "real world knowledge" is another option for disambiguation of ambiguous texts. In this case, knowledge of the real world is employed to rule out some potentially ambiguous texts. Take the following well-known example:

The man saw the girl with the telescope.

Real world knowledge is not sufficient to disambiguate this sentence as both "man" and "girl" can hold a telescope. We would therefore have to resort to contextual information, with all its shortcomings, in order to attempt a disambiguation. If the above example is compared, however, with that below:

The man saw the horse with the telescope.[72]

we conclude that real world knowledge could be applied in order to disambiguate this example.

[72]Examples taken from Hutchins & Somers, op cit, 1992.

If the use of any of the above strategies fails to disambiguate sentences, then other strategies can be used, for example, the "best guess" approach where the computer chooses that analysis which seems most plausible or the interactive approach, where the computer asks a human operator questions in order to disambiguate. Finally, there is what is known as the "free ride" strategy which involves retaining the ambiguity in the target language, for example

The man saw the girl with the telescope.

-> *L'homme a vu la jeune fille avec le téléscope.*

or

-> *Der Mann sah das Mädchen mit dem Teleskop.*

It is obvious that the greater the number of ambiguities in a text the more problematic that text will be for automatic translation. It follows then, that the more ambiguous a text is, the less suitable it is for automatic translation. This does not suggest, however, that texts which display ambiguities are totally unsuitable for machine translation. They will simply require a more sophisticated grammar, and even then may be so problematic as to necessitate human intervention. We propose that an analysis of the degree of ambiguity displayed by different text types is a sufficient platform from which to identify text types which are suitable (or more suitable) for machine translation. Such an analysis will also reveal the similarities and differences between various text types in the same sublanguage.

On the basis of this assumption certain linguistic features from the categories of lexical, semantic and syntactic ambiguity are chosen and our corpus of texts is analysed for each of these features. An empirical analysis of the texts for every linguistic feature which causes ambiguity is beyond the scope of this study. Time constraints allowed us to choose a maximum of two features under each of the three headings. This analysis is

sufficient to allow us to demonstrate the differences between the three text types in question. Thus, under the heading of syntactic ambiguity, conjunction and ellipsis are examined. In the category of semantic ambiguity, occurrence of referential pronouns and compound words are examined (as already stated, there is little point in considering translational ambiguities as this is a monolingual study). And, finally, as there are no occurrences of synonyms in our corpora, the two other phenomena which cause semantic ambiguity are examined, i.e. grammatical category and polysemy. In the section on syntactic ambiguity verb forms are also examined. Although they do not cause ambiguity as such, an analysis of the types of verb forms which occur reveals interesting statistics about the text types under examination.

3.3 Analysis

3.3.1 Semantic Ambiguity

3.3.1.1 Reference

Introduction:

Equipped with general knowledge, co-textual and contextual information, the human processor usually has no problem determining the co-referent of a referential pronoun. Resolution of anaphoric pronouns by computers, on the other hand, can be very difficult as the computer does not have access to contextual or to real world knowledge. Resolution of anaphoric pronouns is crucial for machine translation when translating into target languages which mark pronouns for gender. The following example illustrates the problem:

(a) *The monkey ate the banana because it was hungry.*

(b) *The monkey ate the banana because it was ripe.*

(c) *The monkey ate the banana because it was teatime.*

In each of the above examples, the pronoun *it* refers to a different referent: In example (a) *it* refers to *the monkey*, in (b) *it* refers to *the banana* and in (c) *it* refers exophorically to a specific time period. When we attempt to translate these examples into German, for example, the pronoun changes according to each referent's gender:

(a) *Der Affe hat die Banane gefressen, da er Hunger hatte.*

(b) *Der Affe hat die Banane gefressen, da sie reif war.*

(c) *Der Affe hat die Banane gefressen, da es die Teestunde war.*[73]

These examples illustrate how important the correct interpretation of a pronoun is for automatic translation.

Method

The texts in each of our corpora were analysed and the number of pronominal references in each one was recorded. It is necessary to point out that not every pronominal reference is ambiguous. However, the more pronominal references there are, the more potential ambiguities there are. It was decided to analyse our sublanguage texts according to the framework proposed by Halliday and Hasan since this is generally considered the most comprehensive treatment of connectivity in texts and has become the standard textbook on the subject.[74]

According to Halliday and Hasan, there are three types of reference: personal, demonstrative and comparative. Personal reference is expressed through the category of person. Demonstrative reference is expressed by means of identity or similarity. For this analysis, personal reference, including the three categories of personal pronouns, possessive adjectives and possessive pronouns, are considered along with demonstrative reference which includes the nominal demonstratives *this, that, these,* and *those*. Comparative reference is not considered because it simply compares items in terms of likeness and unlikeness using words such as *other, similarly, fewer, additional* etc. which are not usually a source of ambiguity in texts.

---

[73]These examples are from Hutchins & Somers, op cit, p. 95.

[74]Halliday, M.A.K. & R. Hasan, *Cohesion in English*, (London & New York: Longman, 1976).

Results

The results of our analysis of referential pronouns in each corpus are presented hereunder. The pronoun is presented in the left hand column of the table. The second column entitled "Instructions" presents the results for the Sewing Instructions text type. The next column entitled "Envelope" presents the results for the text on the back of the Pattern Envelope. Finally the last column, "Textbook", presents the results from the sample of text from the school Textbook on sewing.

Table 3.2: Nominal Demonstratives

Referential Item	Instructions	Envelope	Textbook
this	0	0	17
that	0	0	0
these	0	0	0
those	0	0	0

Table 3.3: Personal Pronouns

Referential Item	Instructions	Envelope	Textbook
I	0	0	0
me	0	0	0
you	0	0	12
he	0	0	0
she	0	0	0
we	0	0	0
us	0	0	1
it	0	0	25
they	0	0	3
one	0	0	0

Table 3.4: Possessive Pronouns

Referential Items	Instructions	Envelope	Textbook
mine	0	0	0
yours	0	0	0
ours	0	0	0
his	0	0	0
hers	0	0	0
theirs	0	0	0
its	0	0	0

Table 3.5: Possessive Adjectives

Referential Item	Instructions	Envelope	Textbook
my	0	0	0
your	0	0	1
our	0	0	0
his	0	0	0
her	0	0	0
their	0	0	0
its	0	0	2
one's	0	0	0

An examination of the tables above reveals that reference occurs in only one text type of the three under consideration. In both the Sewing Instructions and the Pattern Envelope corpora there are no referential pronouns. With the exception of the possessive pronouns, the Textbook corpus displays every type of pronoun. Why is this the case? We can only surmise, on the basis of our knowledge of these text types, why no referential items occur in the Instructions and the Pattern Envelope. Let us first look at an excerpt of text from the corpus of Sewing Instructions:

Right side uppermost and matching small circles, make tucks at upper edge of right front skirt in directions of arrows indicated on pattern. Tack to hold. Machine neaten edge of attached facing.

This extract is typical of the style and register of Sewing Instructions. In general, instructions must be clear, succinct and logically ordered. In the second last line "Tack to hold" there is ellipsis of what could either be a noun or the pronoun "it". However,

due to the need to be succinct, neither the noun nor the pronoun is used. It must also be considered that these instructions are describing how to carry out a physical (and visual) process, in which the language plays only a secondary role. Moreover, the instructions are accompanied by diagrams which also act as an aid to the sewer. For these reasons referential items are not necessary in this text type.[75]

Let us now examine the Envelope corpus. As previously stated, the texts in this corpus are even more succinct than the Sewing Instructions. This text type consists of a description, e.g.

Dress with gathers from dropped waistline has front button fastening. View 1 with extended shoulders has shoulder yoke, side seam pockets and three-quarter length sleeves with turn back cuff. Worn with purchased belt.

which is followed by lists of measurements and types of suitable fabric. In a list, the use of referential items is not necessary. In the last line of the description given above there is ellipsis either of the noun or the pronoun and the copula "is". Although it is possible to use a pronoun here, i.e. ***It is*** *worn with purchased belt*, it is not necessary and, indeed, considering the relatively small amount of space available on the envelope packet, ellipsis of the pronoun is probably the most convenient choice.

In contrast with the text types mentioned above, the main function of the Textbook is to teach people how to sew. The realisation of this function involves a combination of text types: (i) descriptive passages informing about the uses/advantages/disadvantages of using a particular method and (ii) instructions on how to carry out this method. It seems that referential pronouns are required in this particular text type.

---

[75]Note that in his comparison of features in six sublanguages (Kittredge & Lehrberger, op cit) Kittredge found that certain linking devices, such as pronominalisation, are apparently not used in a number of technical sublanguages. He notes, and this is what is of interest to us, that pronominalisation tends not to be used in those sublanguages "where the purpose of the text is the most rigidly defined" (p.128).

We will now examine the occurrence of pronouns in the Textbook corpus more closely.

Nominal Demonstratives - Textbook

Of the four lexical items, *this, that, these* and *those*, all occur except *these*. All occurrences of *that* and *those* are relative pronouns rather than nominal demonstratives. *This* occurs 17 times as a nominal demonstrative. Upon examination of the occurrences of *this* we note that all occurrences are not identical. In some cases the demonstrative completely replaces a noun:

(a) Topstitching: This outlines or emphasises the cut of a garment.

and, in other cases, it either recurs with a noun which has previously been mentioned or, alternatively, it occurs with a more general noun which is used as a superordinate for the previous noun or, in some cases, the previous sentence. Examples of these phenomena follow:

(b) Machine felled seam: This seam is prepared, worked and finished in the same way as the Run and Fell Seam.

(c) Working from right to left, make a double backstitch and bring 2 mm to left of first stitch. Keep stitches 2 mm in size. Repeat this process leaving no spaces between stitches.

In example (a) the demonstrative is behaving as "Head" of the noun phrase while in examples (b) and (c) it behaves as "Modifier". According to Halliday and Hasan, there is a significant difference between demonstrative as head and demonstrative as modifier as the type of reference is limited depending on the role of the demonstrative:

> A demonstrative as Modifier ("demonstrative adjective") may refer without restriction to any class of noun. A demonstrative as Head ("demonstrative pronoun"), on the other hand, while it can refer freely to non-humans, is highly restricted in its reference to human nouns; it cannot refer to a human referent except in the special environment of an equative clause.[76]

Moreover, if a demonstrative is used as head then the meaning is identical with that of the presupposed item, whereas if it is used as a modifier then reference may be broader referring to the general class denoted by the noun "including but not limited to the particular member or members of that class being referred to in the presupposed item".[77]

The ambiguity generated by this phenomenon becomes apparent when we consider that, for example, the lexical item *this* can be translated, depending on the context and the gender, into several different renditions in French: *ce, cet, cette, ceci, celui-ci, cette-ci* etc. A machine translation grammar would have to be capable of, firstly, identifying the referent in the English source language text and, secondly, rendering the number and gender correctly in the target language.

We have already discussed the fact that the item presupposed by the referent can be either identical or general and that the demonstrative can act as a modifier. One other point needs to be mentioned in conjunction with this: the distance one has to travel through a text in order to find the referent. It is possible that the presupposed item occurs in the same sentence or in the previous sentence. On the other hand, it is also possible that the presupposed item and the presupposing pronoun are separated by several lines of text. There are many examples in our corpus where there is a a heading (e.g. *Saddle Stitch*) followed by three or four points of information about that particular stitch. Frequently *this* appears in the last point but refers back to the heading "Saddle Stitch". This would be problematic for an MT grammar as there are no rules dictating where the presupposed item might be located in the text.

---

[76] Halliday and Hasan, op cit, p.62.

[77] Halliday and Hasan, op cit, p.64.

Moreover, the demonstrative can refer to a sequence of sentences in the text rather than a single noun. This is known as "extended reference" and can only be achieved by the singular forms of *this* and *that* acting as head (and not as modifiers). Halliday and Hasan maintain that

> Extended reference probably accounts for the majority of all instances of demonstratives in all except a few specialized varieties of English.[78]

There are many examples of the demonstrative *this* as extended reference in the corpus:

Begin with a double backstitch and leave threads loose at the end of each row.
This will be used for pulling up gathers.

where *this* actually refers to the "threads hanging loose at the end of each row", something which would again cause problems for automatic translation.

Personal Pronouns - Textbook

Four personal pronouns occur in the Textbook corpus: *you, us, it,* and *they*. *You* occurs 12 times and each time it refers exophorically to the reader of the text, e.g.

If you begin a row of stitching, complete it to keep the tension even.

Us occurs only once and refers to "humans in general":

Sewing machines have made needlework much easier for us.

[78] Halliday & Hasan, op cit, p. 66.

The occurrences of the pronouns *it* and *they* are more interesting from an MT point of view. *It* occurs 25 times in the corpus and all but one of these occurrences are anaphoric, i.e. they refer backwards in the text to a presupposed item. The resolution of anaphoric pronouns is one of the most difficult problems for MT because, as we have already mentioned in relation to the demonstratives, there are no rules about where the presupposed item might be located. It may be at the beginning, middle or end of the preceding sentence. It may be located three or four sentences back or even in the previous paragraph or, similar to extended reference by the demonstratives, *it* may refer to a series of previous sentences. Finally, the pronoun may also refer forward in the text. This is known as "cataphoric" reference. In this particular sublanguage, extended reference using *it* is uncommon. Twenty-four out of twenty-five occurrences of *it* refer simply to a preceding noun, e.g.

Insert needle again at right angles to raw edge. Bring it through to right side forming a diagonal stitch 5 mm deep...

There is one single occurrence of the pronoun *it* which is cataphoric:

Some people find it easier to sew from left to right especially people who are left-handed.

Here *it* refers to the process of "sewing from left to right". This is the only example of extended reference using the pronoun *it*.

They occurs only three times in the corpus. In one case it presents us with the same problem as the anaphoric *it* and the demonstratives:

Never use knots to secure embroidery threads. They ruin the appearance of the finished work.

Knowledge of typical structures of the English language and a specialised knowledge of the subject area allows the reader to interpret "knots" as the referent instead of "embroidery threads" in this example. Somehow this type of knowledge would have to be incorporated into a grammar to secure a correct interpretation.

Conclusion

From our examination of pro-forms in the three corpora we conclude (i) that the Pattern Envelope and Sewing Instructions texts are very similar because neither make use of referential pronouns and that the Textbook text differs in this respect and (ii) due to its use of potentially ambiguous pronouns, the corpus of didactic text would be more problematic for machine translation. This is, however, only an initial impression which might well be proved incorrect by further analysis. Let us now move on and examine the corpora for the next linguistic feature which can cause semantic ambiguity - compound words.

3.3.1.2 Compound Words

Introduction

Noun compounds (also called noun strings, noun sequences or, in the domain of sublanguage research, *empilages*) are defined as "...two or more nouns plus necessary adjectives (and less often verbs and adverbs) that together make up a single concept; that is, the total expresses a "single noun" idea".[79] Compound words can cause problems for machine translation because the grammar must be capable of recognising a group of lexical items as one compound or noun string and must not parse or translate them separately. On the subject of *empilages* Lehrberger states that

> The proper bracketing of an empilage requires an understanding of the semantic/syntactic relations between components.[80]

The example he cites is *main fuel system drain valve* from the sublanguage of aviation hydraulics. Does *main* modify *fuel system* or *drain valve* in this case? Lehrberger states that it is possible to determine the relations between parts of a compound, e.g. a WHOLE-PART relation, and that this may be sufficient for analysis of empilages in the sublanguage under investigation by him. This may, of course, also be true for many other sublanguage domains.

Lehrberger maintains that in the sublanguage of aviation hydraulics, noun sequences result from "the need to give highly descriptive names to parts of the aircraft in terms of their function in the aircraft and their relation to other parts".[81] He goes on to suggest that this is a characteristic of texts "describing very complex machinery

---

[79]Trimble, L., *English for Science and Technology: A Discourse Approach* , (Cambridge: Cambridge University Press, 1985), pp. 130-131.

[80]Kittredge & Lehrberger, op cit, p.92.

[81]Kittredge & Lehrberger, op cit, p.92.

containing a large number of specialized parts". We would argue that the phenomena of a high occurrence of noun strings extends beyond this narrow range of texts to include such texts as sewing instructions and didactic texts on sewing as well as many more sublanguage texts. The need to give "highly descriptive names" to concepts in the sublanguage and to describe them "in terms of their function...and their relation to other parts" exists in the sub-domain of sewing too.

Trimble discusses compounds from the point of view of teaching English for science and technology to non-native speakers. Although his approach is not entirely relevant to this study, his rather arbitrary system for classifying compounds into the four classes of **simple, complex, more complex** and **very complex** provides an adequate framework for the classification of compounds in the text types under examination here. Below are examples of compounds which fall into one of the above categories.

Simple:

metal shaft

Complex:

Liquid storage vessel

Automated nozzle brick grinder

More complex:

Aisle seat speech interference level

Very complex:

Full swivel steerable non-retracting tail wheel overhaul

Compounds are allocated to one of the four categories above according to the length of the string. Hence, a two-word compound is classified as "simple", a three- or four-

word as "complex", a five-word as "more complex" and, finally, anything above five as "very complex".

In our analysis of compounds in the three text types under examination it is not our intention to examine the relations existing between component parts of each compound. That is an exercise which falls outside the scope of this study where we are concerned primarily with an examination and comparison of the *number* and *type* of compounds in each corpus.

Method

The compounds which occur in each of our three corpora were marked manually. Following this, the number of occurrences of each compound was counted using the Wordcruncher system. Lists of the compounds found in each corpus are provided in appendices D, E and F. The following table presents the results of the analysis.

Table 3.6: Compounds I

COMPOUNDS	INSTRUCTIONS	TEXTBOOK	ENVELOPE
Total no. of words in corpus	**5232**	**5274**	**1335**
No. of compounds in corpus	**184**	**153**	**62**
Total no. of occurrences	**564**	**387**	**116**

It can be seen that there is a significant difference between the Sewing Instructions Corpus and the Textbook Corpus. Both have a similar number of words but the Textbook Corpus does not have as many compounds and has a significantly lower rate of occurrence of compounds than the Sewing Instructions Corpus. As the number of words in the "Pattern Envelope" corpus is smaller than the other two corpora it was decided that the results could best be compared in terms of a percentage of the total number of words in the corpus. The number of words which form components of compounds was estimated for each corpus. This number was then converted into a percentage of the total number of words. The results are presented in the table below:

Table 3.7 : Compounds II

CORPUS	% of total no. of words which form compounds
Sewing Instructions	**31.44%**
Textbook	**15.04%**
Envelope	**18.43%**

From this evidence we conclude that there is a greater degree of compounding in the Sewing Instructions text type than in either of the other text types. The didactic text and the Pattern Envelope text have a similar rate of compounding at 15.04% and 18.43% respectively while the Sewing Instructions text type has almost double the rate of occurrence of compounds at 31.73%.

Classification of Compounds

Following Trimble's model, the compounds were classified into the categories of **simple, complex, more complex** and **very complex** according to the number of lexical

components they contain. There are no occurrences of "very complex" compounds. The Sewing Instructions corpus has two "more complex" compounds: *facing centre back seam allowance* and *front skirt side seam allowance*. There are no examples of "more complex" compounds in either of the other two corpora. The Sewing Instructions has 79 complex compounds and 103 simple. The Textbook corpus has 16 complex and 137 simple and, finally, the Pattern Envelope corpus has 6 complex and 56 simple.

If these figures are viewed in terms of the percentage of the total number of compounds the following results are obtained:

Table 3.8 : Compounds III

TYPE	Sewing Instructions	Textbook	Envelope
Simple	55.98%	89.54%	90.32%
Complex	43.93%	10.46%	9.68%
More Complex	1.09%	-	-

The similarities between the didactic text type and the Envelope are once again apparent. Both of these text types have an almost identical percentage of types of compounds. The Sewing Instructions, on the other hand, while having a greater number of compounds, also displays a more even distribution between simple and complex compounds.

Hence, although the Sewing Instructions text type and the Pattern Envelope texts are similar in terms of referential items (neither display any referential items), they are quite different in terms of the degree and type of compounding which occurs. In the compounding category, it is the didactic text and the Envelope text which share similarities and the Sewing Instructions which deviate.

3.3.1.3 Summary

Under the heading of semantic ambiguity the two features reference and compounding have been examined. The examination of reference revealed that the didactic-instructive text type from the Textbook corpus used referential items, making it potentially ambiguous, while the Sewing Instructions and the Pattern Envelope displayed no referential items. When compounding was examined, similarities were observed between the Textbook corpus and the Pattern Envelope. In the Sewing Instructions, however, there was a greater number of compounds, a higher frequency of occurrence of compounds and, furthermore, there was a greater number of "complex" and "more complex" compounds than in the other corpora.

In the next section we will proceed to examine the texts for features which cause syntactic ambiguity beginning with conjunction, followed by ellipsis and, finally, verb forms.

3.3.2 Syntactic Ambiguity

3.3.2.1 Conjunction

Lehrberger and Bourbeau define the concept of "simple sentence" as

> one in which there is no embedding, no conjunction and no modification within a constituent except by determiners or verb auxiliaries: a simple sentence may not contain more than one (non-auxiliary) verb.[82]

They claim that linguists have been fairly successful in writing computer programs for automatic parsing and generation of simple sentences. On the other hand, automatic parsing of complex sentences, where embedding, conjunction, etc. do occur, requires a very sophisticated syntactic component. For example, the arguments of each verb, if there is more than one verb in the sentence, must be identified; the scope of each conjunction must also be determined as the conjoined elements may be verb phrases, adjectival phrases, noun phrases, prepositional phrases and so on. Moreover, deletions associated with conjunction and embedding must also be accounted for. An example quoted by Lehrberger and Bourbeau from an aircraft maintenance manual illustrates the type of problem posed by conjunction:

The function of the priority valve is to restrict fluid flow to the secondary sub-systems ***and*** *to supply fluid on a priority basis for operation of the flight controls.*[83]

Here it is not clear if "and" is a conjunction of predicates meaning "to restrict ... and to supply" or a conjunction of prepositional phrases meaning "to the secondary sub-systems and to supply fluid" where "supply fluid" is a noun-noun compound.

---

[82] Lehrberger, J., & L. Bourbeau, *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation,* (Amsterdam and Philadelphia: John Benjamin's Publishing Co.,1988), p.89.

[83] Lehrberger & Bourbeau, op cit, p.92.

Co-ordinate conjunctions such as *and* and *or* are particularly problematic for MT. This is due to the fact that elements of almost any category can be joined by a co-ordinate conjunction:

> A major difficulty is the recognition of the scope of conjunctions, e.g. whether **und** ("and") coordinates two nouns, or two noun phrases, or two verb phrases, or whatever.[84]

Conjunction not only links elements in a sentence, but also describes the semantic relations between the two linked elements. The scope and complexity of the relation vary from one type of conjunction to the next. Halliday and Hasan (1976) list conjunction as one of the cohesion producing features in a text. According to them, conjunction is such a complex relation that it can only be divided into four very broad categories which express the general semantic relations "additive", "adversative", "temporal" and "causal". These categories are exemplified by the conjuncts *and, yet, then* and *so*. To these four categories a fifth is added: the category of "continuatives", i.e. words which cannot be classified under any of the above semantic relations but which nevertheless contribute in some way to the cohesiveness of a text. Examples of such words are *now, well, anyway, surely* etc.

Chalker provides a more detailed classification of conjunctive items according to the traditional categories of co-ordinating and subordinating conjunctions.[85] Co-ordinating conjunctions join elements that are grammatically equal while subordinating conjunctions join a subordinate or "dependent" clause to a main clause. Examples of co-ordinating conjunctions are *and, or, but, so, yet, for, neither...nor, either...or* and *not only....(but also)*. Subordinating conjunctions are divided into more specific categories of time, place, comparison, cause/reason, purpose/result, condition, concession/contrast, manner, comparison and proportion. The following list presents those conjunctive items which occur in each of these categories:

---

[84] Hutchins & Somers, op cit, p.199.

[85] Chalker, S., *Current English Grammar*, (London: Macmillan Press, 1984).

TIME:

after, as, before, until, when, now (that), immediately, once, till, whenever, while, as/so long as, as soon as, no sooner...than, hardly/scarcely...when

PLACE:

where, wherever

CAUSE/REASON:

as, since, seeing (that), because

PURPOSE/RESULT:

so (that), in case, such (that)

CONDITION:

if, unless, provided (that), providing (that), on condition (that), as/so long as, suppose (that), supposing (that), whether...or, in case

CONCESSION/CONTRAST:

although, though, while, whilst, whereas, even if, even though, adj/adv + as/though, much as, no matter (what/who)

MANNER:

as, as if, as though

COMPARISON:

adv + than (e.g. longer than)[86]

In addition to these categories, the relative pronouns *who, whoever, which* and *that* are classified as "subordinators", or words which introduce subordinate clauses. It should be noted that there can be overlap between categories, e.g. *as* (Manner, Time, Cause/Reason).

Both subordinating and co-ordinating conjunctions may consist of (a) single words (e.g. *and*), (b) compounds (e.g. *as long as*) or (c) correlatives, i.e. conjunctions that split around an adjective or adverb (e.g. *so....that*). Co-ordinating conjunctions are usually positioned at the beginning of the second clause. They may also be found at the beginning of a sentence and, in this case, they co-ordinate two sentences rather than two clauses. Frequently following the co-ordinate conjunctions *and, or* and *but* there is ellipsis of the second subject, e.g. *She left the house in a rage, but ϕ didn't get far.*[87] Subordinate conjunctions also come at the beginning of the clause and, like co-ordinating conjunctions, can come in first position in the sentence. The main semantic difference between the two types of conjunction is that propositions joined by co-ordinating conjunctions have equal status, i.e. they are both offered as new information in the sentence. The content of a subordinate clause, on the other hand, is presented as "given". A subordinate clause, therefore, downgrades its information and helps to emphasise the main clause.[88]

---

[86]Chalker, S., op cit, pp.244-249.

[87]The symbol ϕ indicates the position where ellipsis occurs in the sentence.

[88]Chalker, S., op cit, p.239.

METHOD

Following Chalker's classification system, each corpus was examined for both co-ordinating and subordinating conjunctions. The results were compiled separately for each corpus. They were then compared and contrasted. Finally, the texts were examined in more detail for the most common conjunctions and conclusions were drawn from this analysis about syntactic ambiguity caused by conjunction in each text type.

CONJUNCTION IN PATTERN ENVELOPE CORPUS

There are no examples of subordinating conjunctions or relative pronouns which introduce subordinate clauses in this corpus. Co-ordinating conjunctions do occur. However, there are only two: *and* (29 occurrences) and *or* (25 occurrences).

CONJUNCTION IN SEWING INSTRUCTIONS CORPUS

Like the Pattern Envelope corpus, the Sewing Instructions corpus only displays two co-ordinating conjunctions: *and* (163 occurrences) and *or* (6 occurrences). Unlike the Pattern Envelope, however, the Sewing Instructions display a range of subordinating conjunctions as illustrated in the table below:

Table 3.9: Subordinating Conjunctions - Sewing Instructions

Type	Conjunction	Number
Time	before	2
	when	1
Place	where	9
Condition	if	2
Manner	as	75
Comparison	as...as...	4

It is clear from the table above that the highest occurring subordinating conjunctions are *where* and *as*. All occurrences of *where* read *where necessary* in the corpus. With the exception of the comparative *as*, which in each instance read *as far as possible*, all other occurrences of *as* are *as shown*.

CONJUNCTION IN TEXTBOOK CORPUS

In comparison with the other corpora, this corpus of texts displays the widest range of both subordinating and co-ordinating conjunctions:

Table 3.10: Co-ordinating Conjunctions - Textbook

Conjunction	Number
and	153
or	50
but	5

While the number of occurrences of the conjunction *and* is similar to that of the Sewing Instructions corpus (163), the number of occurrences of *or* is significantly higher in the Textbook corpus (50 vs. 6). Furthermore, this is the only corpus where the conjunction *but* occurs.

Table 3.11: Subordinating Conjunctions (TIME) - Textbook

Conjunction	Number
after	2
as	4
before	6
until	7
when	11
now	2

Table 3.12: Subordinating Conjunctions (Other) - Textbook

Type	Conjunction	Number
Place	where	4
Cause/Reason	as	4
Purpose/Result	so that	4
Condition	if	6
Concession/Contrast	however	1
	no matter	1
Manner	as	7
Comparison	...than	4

While subordinating conjunctions do not occur in the Pattern Envelope corpus, they do occur in the other two. The range of semantic relations (time, manner, cause etc.) covered by subordinating conjunctions is greater in the Textbook corpus (8) than in the Sewing Instructions corpus (5) (see tables above). This would suggest that the subordinating conjunctions in the didactic text type (i.e. the Textbook corpus) covers a wider semantic area than those of the Sewing Instructions and that, in turn, the latter covers a wider semantic area than the Pattern Envelope text type.

The degree of co-ordination is similar in both the Sewing Instructions corpus and the Textbook corpus. In terms of a percentage of the total number of words in each corpus, the three corpora are similar: the Pattern Envelope corpus has 4.02% coordinating conjunctions, the Sewing Instructions has 3.23% and the Textbook corpus has 4.15%.

It was mentioned previously that the co-ordinating conjunctions *and* and *or* are most problematic for MT. Both *and* and *or* occur in each of the three corpora. The occurrences in each corpus are examined below with a view to establishing the extent of ambiguity caused.

Pattern Envelope Corpus

The co-ordinate conjunction *and* causes ambiguities in five out of twenty-nine cases (i.e. 17.24%) in this corpus. For example:

(a) *View 2 is sleeveless with wide neckline and patch pockets.*

In this example it is not clear if the adjective *wide* modifies *neckline* only or both *neckline* and *patch pockets*.

(b) *Allow extra fabric for matching plaids and stripes.*

Example (b) illustrates the same problem with adjective scope as example (a), i.e. does *matching* modify both *plaids* and *stripes*?

(c) *with mock vent and button trim*

In this example it is possible that the lexical item *mock* modifies the compound *vent and button trim*. On the other hand, *mock vent* and *button trim* could be interpreted as two separate compounds joined by a conjunction.

The conjunction *or* presents a similar problem:

(d) *matching plaids or stripes*

where, again, it is not clear whether *matching* modifies both *plaids* and *stripes* or *plaids* alone.

Similar problems occur in the Sewing Instructions and Textbook corpora but the number of ambiguities is relatively small when compared with the Pattern Envelope corpus:

Sewing Instructions

There were only six cases of ambiguity out of a total of 163 occurrences of *and* (i.e. 3.68%) in this corpus.

(e) *With waistband extension to right back skirt and matching large centre front circles, and small circles to side seams, stitch waistband to skirt.*

=> "With waistband extension to right back skirt and **to the matching centre front circles**...?

or

=> "With waistband extension to right back skirt and **while matching** large centre front circles...?

(f) *Stitch front and neck edges.*

=> "Stitch the **front piece** and the **neck edges**"?

or

=> "Stitch the **front edges** and the **neck edges**?

Textbook

The Textbook corpus had only 4 examples of ambiguity out of a total of 153 occurrences of *and* (i.e. 2.61%).

(g) *Also worked on armhole and pocket seams.*

=> "...armhole **seams** and pocket seams"?

(h) *Secure thread ends with a backstitch or a few running stitches along the design and cover with embroidery stitches.*

=> "...along the design and **cover**...": Verb?

or

=> "along the design and **cover**...": Noun?

It was noted that the conjunction *or* did not cause any ambiguity in either the Sewing Instructions or the Textbook corpora while there were three ambiguous examples in the Pattern Envelope corpus.

The ambiguities caused by co-ordinating conjunctions in our corpora are, for the most part, due to one of three reasons. The category of one of the lexical items may be ambiguous as in (e), (f) or (h) above. For example, in (f) the lexical item *front* may be a head noun or a modifier. Secondly, the scope of the adjective may be ambiguous as in example (d) where *matching* may modify both "plaids" and "stripes" or "plaids" alone. Finally, the ambiguity may be due to a lack of syntactic information which allows for proper bracketing of a phrase, as in example (c).

Summary

In summary, the percentage of co-ordinating conjunctions, in terms of the total number of words in each corpus, is similar in all three corpora. The Textbook corpus has the widest range of subordinating conjunctions which cover a wider semantic range than both the Sewing Instructions and the Pattern Envelope (which has no subordination). Of the three text types, the Pattern Envelope text has the highest number of ambiguities caused by conjunction.

Although *and* is the conjunction with the highest number of occurrences in each corpus, the resulting number of ambiguities is relatively small as can be seen from the figures quoted above. This is obviously a plus for the suitability of these texts for automatic translation. In addition, if the grammar of this sublanguage were to be examined with the aim of designing a MT system, disambiguation of at least some of the above examples would be possible. Such an examination would reveal, for example, that the lexical item *matching* in (e) above is a verb and not a modifier. *Matching* does occur in modifier position in the Sewing Instructions corpus but only

ever in the phrase *Fuse to wrong side of each matching fabric section.* All other occurrences of this lexical item are in verbal position in the sentence. Moreover, having established that this lexical item is potentially ambiguous, the authors of these texts could "control" the language by adding the necessary lexical items for disambiguation as we have done in example (e) above, i.e. *to the matching centre front circles* if it is a noun, or *while matching the centre front circles* if it is a verb. Although this would make the language of Sewing Instructions less succinct, significant advantages would be gained by making the sublanguage more suitable for MT through the elimination of potential ambiguities.

3.3.2.2 Ellipsis

Introduction

Ellipsis occurs when part of a sentence, for example the definite or indefinite article or the direct object, has been omitted for reasons of economy, emphasis or style. The omitted element is "understood" by the reader and is usually recoverable from the context.[89] Ellipsis often results in ambiguity and can, therefore, cause problems for machine translation. For example, whereas English uses no article for singular mass nouns in object position, French uses the "article partitif".[90] Hence, a deleted article in an English text will have to be restored during the translation process into French or German. Ellipsis is one of the immediately obvious characteristics of the text types in our corpora. An examination of a sample of each text type is undertaken here in order to obtain an overview of the patterns of ellipsis in each.

Samples of 500 words from both the Sewing Instructions corpus and the Textbook corpus were examined for ellipsis. This involved restoring any obvious deletions of definite or indefinite articles, prepositions, conjunctions etc. and then counting the number of restored items. The number of ellipses that occurred in the Sewing Instructions sample was significantly higher - 160 in a 500 word sample - than the Textbook which had only 71 examples of ellipsis in a 500 word sample of text. Since the Pattern Envelope corpus amounts to only 1335 words in total, it was an easier task to examine the whole corpus for ellipsis than the other two corpora with approximately 5000 words each. From its total of 1335 words, the Pattern Envelope text had 91 examples of ellipsis. This number is slightly higher than the Textbook corpus sample, which is to be expected as there is a greater number of words under consideration, but is lower than the Sewing Instructions corpus even though there is a smaller number of

[89] Crystal, D., *A Dictionary of Linguistics and Phonetics*, (Oxford, Basil Blackwell, 1985), p.107.
[90] van Eynde, F., (ed.), *Linguistic Issues in Machine Translation*, (London & New York: Printer Publishers, 1993), p.10.

words in the Sewing Instructions sample. This indicates that the Sewing Instructions text is more elliptical than the other two. The following table shows a breakdown of the types of ellipsis and the number of occurrences in each sample:

Table 3.13 : Ellipsis

ELLIPSIS OF	SEWING INSTRUCTIONS	TEXT BOOK	PATTERN ENVELOPE
Article	117	56	37
Direct Object	29	3	-
Copula	7	2	-
Prepositional Phrase	7	1	11
Noun Phrase	-	2	-
Verb Phrase	-	10	12
Conjunction	-	-	1

Subject ellipsis

In the text types under consideration, ellipsis of the subject occurs before every imperative verb. This phenomenon, along with other types of ellipsis, is characteristic of instructional text types. Restoring the subject where it has been elided would interfere with the reading and interpretation of the text, e.g. ***You*** *press the material,* ***You*** *stay-stitch the hems,* ***You*** *hold the right sides uppermost* etc. For this reason, the number of elided subjects before imperative verbs are not counted. Restoring other phenomena such as the definite article, the direct object, etc. where they have been deleted could reduce ambiguity. For example, the ambiguous *matching circles* (MOD

+ N or VB + N?), discussed previously under the heading of conjunction, would not be problematic if the definite article were restored, i.e. *the matching circles*. If ellipsis of the subject were to occur in any constructions other than before an imperative verb it would be considered. However, there are no examples of subject ellipsis in any of the samples other than before an imperative verb.

As previously mentioned, the Sewing Instructions has a higher tendency than the other texts to elide lexical items. This phenomenon can be explained by the fact that the language of instructions tends to be "telegraphic", while descriptive language, which forms part of the Textbook text type, is not as telegraphic in nature. Furthermore, given that the Textbook text has a didactic function, it is not surprising that the ambiguities caused by ellipsis are avoided. Ellipsis of the article (mainly definite article with some examples of deleted indefinite articles) is the most common type in all three corpora. In the didactic text type there is a tendency to elide both noun phrases (or parts thereof) and verb phrases while in the sample of Sewing Instructions only the copula is elided. The high dependency on verb forms in the Sewing Instructions may explain why there is no ellipsis of lexical verbs.[91] The Pattern Envelope text has the highest rate of ellipsis of prepositional phrases and, like the Textbook corpus, has a tendency to elide verb phrases. This latter phenomenon does not seem to be characteristic of the Sewing Instructions text. Examples of each type of ellipsis follow:

Article ellipsis:

(i) *Machine neaten ϕ (the) edge of ϕ (the) attached facing.*[92]

(ii) *Stitch ϕ (the) centre back seam.*

[91] See next Section on "Verb Forms".

[92] The symbol ϕ indicates where ellipsis takes place and the elided item is given in brackets, e.g (the).

Ellipsis of direct object:

(i) *Press ϕ (the material).*

(ii) *Tack ϕ (the material) to hold ϕ (it).*

Copula ellipsis:

(i) *...between small circles as ϕ (is) shown.*

(ii) *...top-stitch 6mm from upper edge as ϕ (is) shown.*

Ellipsis of prepositional phrase:

(i) *ϕ (With the) Right side uppermost...*

(ii) *ϕ (With the) Right sides together...*

Noun phrase (or part of NP) ellipsis:

(i) *The colour and fibre in the thread should match those ϕ (colours and fibres) of the garment fabric.*

(ii) *To hold matching stripes...of two garment pieces together for first fitting and machine ϕ (stitching) and hand stitching.*

Verb phrase ellipsis:

(i) *It should be used when ϕ (you are) hand sewing.*

(ii) *ϕ(This stitch is used) To attach interfacing to garment piece...*

Ellipsis of conjunction:

(i) *...above mid-knee, ϕ (and) has raised waist with stitched facings.*

It is clear from this analysis of ellipsis that an MT grammar for English sewing instructions would have to cater for a high degree of ellipsis of articles when translating into an article dependent language such as French. While a bilingual analysis of this sublanguage is beyond the scope of this study, we felt it was useful to briefly examine a parallel sample of this sublanguage in French to establish how problematic ellipsis of the article might be. Parallel garment descriptions from a Pattern Envelope were examined. The initial conclusions from this examination suggest that (i) the French sublanguage is more elliptical than a general language French text; (ii) where the definite article is elided in English it is often necessary to replace it in French, as the following examples illustrate:

(a) *Dresses have front button fastening and flattering panel seams which flare at hem.*

Robes décolletées et butonnage sur ***le*** *devant, panneaux piqués amincissants évasés dans* ***le*** *bras.*

(b) *Model 1 has wide shoulder straps which form a tuck at upper edge of dress.*

Modèle 1 à bretelles larges formants un pli au bord supérieur de ***la*** *robe.*

(c) *...use nap requirements and nap layouts.*

...utiliser ***le*** *métrage et* ***le*** *plan de coupe "avec sens".*

A third and final observation from this brief examination is that verb phrases in the English text are often transposed into noun phrases in the French text, e.g. "has wide shoulder straps" -> "à bretelles larges" or "have front button fastening" -> "buttonage sur le devant". A MT grammar must be able to deal with these types of transpositions in order to adequately translate this sublanguage from English into French.

Summary

We have noted that the Sewing Instructions has a higher degree of ellipsis than the other two text types. When expressed in terms of a percentage of the total number of words in each sample, the following figures are obtained:

Sewing Instructions:	32%
Textbook:	14.2%
Pattern Envelope:	6.8%

The instructive function of the Sewing Instructions text type, which requires telegraphic, succinct sentences, is responsible for the higher degree of ellipsis. While the Textbook text type is also instructive in function, it has a descriptive element which explains why the occurrence of ellipsis is lower here. Finally, the Pattern Envelope text type, which is the most telegraphic of the three, has the lowest occurrence. How can this be so? The text is so telegraphic that it consists mainly of phrases and there are few complete sentences. It would obviously not be wise to have a high degree of ellipsis in the already telegraphic phrases. Hence, the lower occurrence of elided items in the Pattern Envelope text type.

3.3.2.3 Verb Forms

Introduction

Although verbs are not listed as one of the causes of ambiguity, an analysis of the types of verb forms which occur in each corpus provides more information about the linguistic characteristics of each text type. The distribution of verb forms across each text type is illustrated below.

Sewing Instructions

One of the most outstanding features of this text type is the very high use of the imperative verb form and the absence of any past tense. Also of significance is the low occurrence of the present simple tense and of the future tense. The imperative form accounts for 68% of the total verb forms, while the present simple (active) accounts for 0.1% and the future for 0.3%. After the imperative, the next highest occurring verb form is the present progressive at 20%. The only other verb forms which occur are the infinitive (3.3%) and the present simple (passive) (8.3%). The distribution of verb forms across this corpus of texts is summarised in the table below:

Table 3.14 : Verb Forms - Sewing Instructions

Verb Forms	% of Total Verb Forms
Imperative	68%
Present Simple - Active	0.1%
Present Simple - Passive	8.3%
Present Progressive	20%
Future	0.3%
Infinitive	3.3%

Text Book

Similar to the Sewing Instructions corpus, the highest occurring verb form in the Textbook corpus is the imperative. In the latter, the imperative accounts for a smaller percentage of the total verb forms (52.69%) than in the Sewing Instructions corpus. The next highest occurring verb form is the present progressive (Active + Passive) at 17.79%. If the range of verb forms in the Textbook corpus is compared with the Sewing Instructions it becomes apparent that the former uses a wider selection of verbs than the latter including, for example, modals and the present perfect. The following table presents a summary of verb forms in the Textbook corpus:

Table 3.15 : Verb Forms - Text Book

Verb Form	% of Total Verb Forms
Imperative	52.69%
Present Progressive - Active	17.38%
Present Progressive - Passive	0.41%
Present Simple - Active	8.41%
Present Simple - Passive	7.17%
Future	1.10%
Modals - Active Progressive	3.59%
Modals - Active Simple	0.28%
Present Perfect - Simple	0.28%
Present Perfect - Passive	0.55%
Infinitive	8.14%

Pattern Envelope Corpus

This corpus is similar to the Sewing Instructions corpus in that past tense does not occur. Unlike the latter, however, modals are used and future tense is not. Also, the distribution of verb forms in this text type is somewhat different: imperative accounts for only 44% of total verb forms (this figure is lower than both the Sewing Instructions and Textbook corpora) and the next highest occurring form is present simple and not present progressive as in the latter two.

Table 3.16 : Verb Forms: Pattern Envelope

Verb Form	% of total no. of verbs
Imperative	44%
Present progressive - Active	12%
Present Simple - Active	32%
Present Simple - Passive	4%
Modals	8%

Conclusion

As might be expected of an "instructive" text type, the imperative is the most frequently occurring verb form across our three corpora. This is followed by the present progressive in the case of the Sewing Instructions text type and the Textbook and by the present simple in the Pattern Envelope. Another similarity between the three is that there is no occurrence of the conditional tense whatsoever.

Nevertheless, a closer examination reveals considerable differences between the three text types. The text type in the Textbook corpus has a wider range of verb forms than either of the other two. This may possibly result from the texts' broader communicative function(s). Furthermore, while the second most dominant verb form in the Sewing Instructions and the Textbook is the present progressive, the present simple is more dominant in the Pattern Envelope corpus. In general, the range of verb forms in the Pattern Envelope corpus is similar to that of the Sewing Instructions. The imperative, present simple and present progressive are used in both. The differences between these two text types are that modals are used in the Pattern Envelope and not in the Sewing Instructions, while the latter makes use of the future tense and infinitive, both of which are absent from the Pattern Envelope corpus. We conclude, therefore, that the distribution of verb forms across the three text types differs significantly.

3.3.3 Lexical Ambiguity

3.3.3.1 Categorical Restrictions

Introduction

In language as a whole many lexical units can occur in more than one category, for example, *play* can be used as a verb (*The children play in the fields*) or as a noun (*We went to see a play last night)*. Since lexical units can occur in more than one grammatical category, machine translation systems must be capable of recognising and correctly translating the grammatical category of each lexical unit. In sublanguage, it is often the case that a lexical item will occur in only one grammatical category or, alternatively, a lexical unit which usually occurs in one category in standard language (e.g. a noun) may occur in a so-called "deviant" form (e.g. as a verb) in the sublanguage (For example, in the sublanguage of Sewing Instructions the lexical item *machine* is used as a verb rather than a noun, which is its most common category in standard language). This results in a reduction in the amount of processing to be carried out by the MT system as well as a reduction in ambiguities. The number of multi-categorial items is therefore a significant factor in assessing the suitability of a text for MT.

Method

All unique lexical items in the three corpora were observed and each one that occurs in more than one category was noted. A comparison is made below between the number of multi-categorial items in each corpus. The categories are classified according to their grammatical function, e.g. verb, noun, adjective etc. Many nouns (and some verbs) occur as modifiers of noun phrases where one noun (or verb) modifies another

noun. In some MT systems, e.g. Eurotra, such nouns or verbs are labelled as "modifier" to differentiate them from the head noun. There are many examples of this phenomenon in our corpora and we felt it was wise to follow the practice of labelling such nouns and verbs to differentiate them from head nouns. In fact, all multi-categorial items in the Pattern Envelope text type are nouns and modifying nouns (with the exception of one, c.f. Appendix L). This is due, we believe, to the highly nominalised nature of this particular text type where the information load is carried by nouns and their modifiers rather than verbs. Hence, the label "noun as modifier" (or N (mod)) is used for some lexical items which, although they do not have a different grammatical category, they do have a different grammatical *function*. A list of the multi-categorial items in each corpus is to be found in Appendices J, K and L. The following table shows a comparison of the number of multi-categorial items in each corpus:

Table 3.17: Multi-categorial Items

Corpus	No. of multi-categorial items	No. of unique words in corpus	% of total no. of words
Sewing Instructions	**9**	**418**	**2.15%**
Textbook	**64**	**800**	**8.13%**
Pattern Envelope	**11**	**342**	**3.22%**

From this table we observe that the Textbook corpus has the highest number of lexical items in more than one category, followed by the Pattern Envelope corpus and, finally, by the Sewing Instructions corpus. All multi-categorial items occurred in two categories only, with the exception of the item *open* which occurred in three, i.e.

adverb, adjective, and verb. An examination of the categories reveals that the Noun/Verb category is the most predominant followed by N/N (mod) (i.e. head noun/modifying noun) and V/V (mod) (verb/modifying verb). Examples of these multi-categorial words are provided below.

NOUN/VERB:

Hand baste along roll line and seamlines. (V)

Shaping lapel over hand, padstitch lightly front roll line to seamline. (N)

Layer seams. (V)

Sew hem in place, easing in fullness if necessary, and catching in one layer of pocket. (N)

NOUN/NOUN (MOD):

Secure thread ends with a backstitch or a few running stitches along the design and cover with embroidery stitches. (N)

Work from left to right along design line. (N(mod))

Twilled Cotton. (N)

Cotton Sateen. (N(mod))

VERB/VERB (MOD)

Used for seams, tucks, gathering, easing and repairing garments. (V)

Size of easing stitch = 2-3 mm. (V(mod))

Run thread end through fold and cut. (V)

Run and fell seam. (V(mod))

Conclusion

Yet again we note that the three text types are not linguistically identical. In this case, the Sewing Instructions is most restricted while the Textbook text type is least restricted. Although there are items which have more than one category in all corpora, the percentages in the Sewing Instructions and Pattern Envelope are low. Furthermore, any ambiguities which might occur as a result of these multi-categorial items could be resolved by the syntactic component of a MT grammar which could distinguish a verb from a noun by virtue of its sentential position, for example, the verb *layer* always occurs at the beginning of a sentence while the noun *layer* occurs in the middle of the sentence. Similarly, the syntactic component of the grammar should be able to distinguish a noun (or verb) from the corresponding noun or verb acting as a modifier by virtue of the fact that the modifying noun will normally precede another noun and, similarly, the modifying verb will precede a noun (c.f. the *design/design line* example above).

3.3.3.2 Polysemy

Introduction

According to Leech, the term "polysemy" signifies "the existence of more than one semantic specification for the same lexical item".[93] Leech also discusses the debate about the difference between homonymy and polysemy. However, as already mentioned in the introduction to this chapter, the difference between the two is of little consequence to MT as they are both treated in the same manner.

When polysemy is discussed in linguistic circles it is usually with reference to words such as *bank* or *bridge* which have more than one distinct meaning. The sublanguage approach to MT assumes that, in a restricted domain, there will be few polysemes and the task of translating from the source text to the target text will be made easier as a result. Our aim is to establish the extent to which polysemy occurs in the three text types under consideration.

Method

As our analysis of categorial restrictions has indicated, a significant number of lexical items can occur in more than one category, even in a highly restricted sublanguage. We are now faced with the problem of deciding whether lexical items such as *stitching*, which functions as both a noun and a verb in our corpora, are polysemous. When compared with lexical items such as *bridge* or *play*, which have distinct meanings in different contexts, *stitching* would not seem to adhere to the definition of a polysemous item. However, Hutchins points out that words such as *green* (ADJ + N) and *control* (V + N) are often treated as homographs by MT systems "for practical

[93] Leech, G., *Semantics*, (Penguin, 1974), p.230.

purposes".[94] This would suggest that lexical items which occur in more than one grammatical category should be treated as polysemes. Furthermore, one of the most common methods for differentiating polysemes in MT is to allocate "semantic features" to each one and compare the results. When we apply this process to *stitching* as a verb and *stitching* as a noun we see that there are significant differences in meaning between the two: *stitching* as a verb refers to a process and requires an agent or a "doer" which may be either a human or a machine; *stitching* as a noun, on the other hand, refers to the finished product of a sequence of stitches. We conclude, therefore, that it is necessary to consider the multi-categorial items in our corpus for polysemy.

Before going on to consider the occurrence of polysemous items in each text type, it is necessary to consider the special category of N/N (mod) and V/ V (mod). It is frequently the case in the sublanguage of sewing instructions that N(mod) expresses the relation "is a part of" or "is used to" with its corresponding noun. So, for example, *design* signifies a whole while *design line* signifies "part of" the whole. Similarly, *embroidery* signifies a process while *embroidery thread* signifies something that "is used for" this process. It is apparent from these examples that in some cases the meaning of the modifying noun (N (mod)) is determined by the noun it modifies and not by the modifier itself, i.e. "design" in "design line" depends on its co-occurrence with "line" for its meaning. Thus, we cannot consider lexical items which occur as modifiers of nouns and verbs as polysemes of the corresponding head noun or verb. There are, however, some exceptions to this where the modifier expresses a different meaning to that of the head noun or verb (See example *back* further on). These will be included in the numbers of polysemes in each corpus.

[94]Hutchins, W.J., *Machine Translation: Past, Present, Future*, (Chichester & New York: Halstead Press, 1986), p.42.

Analysis

The following table shows the number of polysemous items in each corpus and the percentage of the total number of words in each corpus:

Table 3.18: Polysemy

Text Type	No. of Polysemes	%
Sewing Instructions	**5**	**0.095%**
Textbook	**32**	**0.606%**
Pattern Envelope	**2**	**0.149%**

Our table reveals that the occurrence of polysemous items in all corpora is low, not reaching 1% of the total number of words in any case. If our corpora are exemplary of this particular sublanguage, then it is reasonable to suggest that the results are encouraging for MT. A low occurrence of polysemy means fewer problems when parsing the sublanguage.

It should be noted that the didactic text type has a higher occurrence of polysemous items than the other two text types. We will now look at some examples of those polysemes more closely.

Sewing Instructions

The five polysemous items in the Sewing Instructions corpus are: *back, ease, finishing, fold* and *layer*. In keeping with the definition of polysemy above, they are all examples

of lexical items with different, but nevertheless related semantic specifications as the following examples illustrate:

Back

*(a) Pin right front to **back** at right side seam.*

*(b) Work a few **back** stitches at circles.*

*(c) Turn **back** neatened edge.*

where (a) refers to a part of the garment, (b) to a type of stitch and (c) to an action.

Ease

*(a) **Ease** in fullness where necessary.*

*(b) A pleat will form at lower edge for wearing **ease**.*

*(c) Adjust **ease**, distributing evenly between notches.*

where (a) is the action of moving with careful manipulation, (b) signifies freedom from discomfort and (c) refers to a piece of material which prevents discomfort.

Finishing

*(a) Stitch front and neck edge starting and **finishing** at front panel seams as shown.*

(b) *Tie End and* ***Finishing***

where (a) refers to an action and (b) is part of a title and refers to the process of sewing a hem, zips, buttons etc. on a garment in order to give it a "finished" look.

Textbook

Some examples of polysemous items from the Textbook corpus follow:

Beginning

(a) *Backstitching is used (i) to secure thread ends at the* ***beginning*** *and end of a row of stitching.*

In example (a) the word *beginning* refers to the start of a row of stitching.

(b) ***Beginning*** *at wide end of dart near edge, machine from edge to point.*

In this example, it refers to the commencement of an action.

Back

Back is one of those lexical items which occurs as a N, a N (mod) (and an ADV) where the N (mod) has a different meaning from the Noun.

(a) *Begin permanent stitches with one or two* ***back*** *stitches.*

(b) *This is used (i) to mark the position of the centre front and centre* ***back*** *of garment.*

(c) *Working from right to left insert the needle through* ***back*** *of fabric.*

(d) *Secure thread, fold* ***back*** *edge of hem.*

(e) *Place the needle through fabric at right side of fabric stitch and bring it out on the left, half way* ***back*** *beside first stitch.*

In example (a) (N (mod)) *back* refers to a type of stitch. In (b) (N) it refers to a part of the garment while in (c) (N) it signifies the "reverse side of" the fabric. Finally in (d) (ADV) it modifies the action *fold* and in (e) the phrase *half way back* modifies the verb *place*. Similarly, in the Sewing Instructions corpus the lexical item *back* can have three of the above readings, i.e. (a), (b) and (d).

The lexical item *right* presents a similar problem in the Textbook corpus. It can occur with the following meanings:

(a) *Select the right type of needle and thread.*

(b) *Place a thimble on the middle finger of the right hand.*

(c) *Some people find it easier to sew from left to right.*

In (a) the meaning is "correct", in (b) it modifies *hand* and refers to a specific part of the body and in (c) it refers to direction.

Pattern Envelope

There are only two polysemes in the Pattern Envelope corpus, i.e. *side* and *back*:

Side

(a) *Side seam pockets*

(b) *Wrong side may show*

where (a) refers to a position or location on the garment and (b) refers to the surface of the material.

Back

(a) *Lining (back/pocket)*

(b) *Turn back cuff*

Here (a) signifies a part of the garment while (b) refers to a style of cuff on the garment.

Conclusion

We conclude from our examination of polysemy that there are few cases of polysemy in these corpora, which is encouraging from the MT viewpoint. The Textbook corpus has most occurrences (32) and would therefore be more problematic for MT than either of the other two text types. Again, we would attribute this fact to the differences in text type between the Textbook and the other two text types. Many multiple meanings in these corpora result from the fact that lexical items can occur in more than one grammatical category. In a MT grammar, the syntactic component should be able to disambiguate many of these occurrences by classifying them according to the correct

grammatical function. For example, the lexical item *ease* has three meanings in the Sewing Instructions. How would these three meanings be differentiated from each other? A syntactic analysis allows us to trace a pattern for each meaning: *ease* meaning "the action of moving with careful manipulation" is an imperative verb which occurs at the beginning of a sentence; *ease* meaning "freedom from discomfort" always occurs in the phrase *for wearing ease*; and, finally, *ease* meaning "a piece of material which prevents discomfort" is always the object of a verb in our corpus. From a sample of approximately 5000 words it is impossible to say if these patterns remain unchanged throughout the sublanguage of Sewing Instructions, although there is little likelihood that they would change. Nevertheless, this example serves to illustrate how an analysis of syntactic features can disambiguate polysemes in a sublanguage.

That concludes our analysis of the three corpora for lexical, semantic and syntactic ambiguity. In the next chapter, the results obtained from our analysis will be evaluated and conclusions will be drawn about the three text types under analysis, about the sublanguage of sewing, in general, and about the advantages of a text-based approach to the analysis of sublanguage texts for MT.

CHAPTER 4 :

CONCLUSIONS:

4.0 Summary

In Chapter 1 an outline of the theories of sublanguage which have been formulated to date was presented. Each of these theories touch on different aspects of sublanguage, for example, the relationship between sublanguage and natural language and general language, the evolution of sublanguage, or the relationship between users of sublanguages. The conclusions drawn from this review are that there is no agreed definition of the term *sublanguage* but there is, nevertheless, a general consensus about the characteristics exhibited by sublanguages. Furthermore, it is also generally agreed that restricted sublanguages are more suitable for machine translation than "standard" or "general" language, thanks to their restricted lexicon and grammar. In the same chapter attention was drawn to the fact that although analysis of texts and text types has been going on in other branches of linguistics, such as translation studies or LSP analysis, since the 1960s, little attention has been given to the text as a unit in sublanguage and MT research. This was attributed to the facts that the theory of *text types* was developing at the same time as the theory of *sublanguage* and that in the domain of MT, researchers were still concentrating on the best *method* of automatic translation and on developing MT systems that could translate successfully so as to prove that MT was a viable domain of research. The results obtained from the analysis of text types in other branches of linguistics suggest that a closer examination of text type in sublanguage and MT research would yield some profitable results. After all, although sublanguages are deemed more suitable for MT than general language texts, many problems still exist, for example, ambiguities caused by long compounds, by conjunction or by anaphoric pronouns. If textual analysis is to become the norm in sublanguage analysis, attention must be paid to the notes of warning sounded by many text type researchers in the past, namely, that it is not an easy task to classify texts according to function because a text can have more than one function, that a text may consist of more than one text type each of which is characterised by different linguistic

features, and that a text's linguistic characteristics are also determined by extra-linguistic features such as communicative situation, knowledge of recipients, author's intention etc. Analysis of sublanguage texts will have to take these factors into consideration.

In order to examine how advantageous a sublanguage text type analysis is, we chose a restricted sublanguage, i.e. that of Sewing, and what appears to be three distinct text types from that domain, for examination. The aims of the analysis were as follows:

(i) *To illustrate that within one sublanguage there is frequently more than one text type and that each text type differs from the next according to its function and other extra-linguistic criteria, and to demonstrate that a text type/text function analysis would be beneficial for MT.*

and

(ii) *To illustrate that although we can talk of "text types" and "text function" there is no one rigidly defined "instructive" text type. Instructive text types can differ linguistically from each other due to the influence of extra-linguistic criteria such as communicative situation, knowledge level of recipients etc. It would therefore be wise to include an analysis of extra-linguistic criteria in any text type analysis.*

Before the analysis could proceed, it was necessary to discuss some of the problems associated with analysing corpora of texts, and, in particular, sublanguage corpora . This topic is discussed in Chapter 2 and on the basis of the discussion guidelines are offered for the compilation of corpora for sublanguage research. Following the recommendations put forward by Hoffmann, to name but one, the functions and extra-linguistic features of our three corpora of texts are described. It was established that two of the text types (i.e. the Sewing Instructions and the Pattern Envelope) were written by an expert for an expert, while the third text type (i.e. the Textbook) was written by an expert for a non-expert. Moreover, while the Textbook text type has primarily a didactic function, it is possible to say that all three text types have a dominant *instructive* function, although they differ significantly in macrostructure.

Having established these general similarities and dissimilarities between the texts, they are then analysed in Chapter 3. The analysis is based on linguistic phenomena which cause ambiguity and which frequently cause problems for MT. Firstly, a description of these phenomena is given. Next, some phenomena from each of the categories of semantic, syntactic and lexical ambiguity are chosen and the texts are subjected to an analysis of each of these phenomena. The current chapter sums up the results of those analyses and draws conclusions based on our findings.

4.1 Semantic Ambiguity: Conclusions

In this category, reference and compounding were examined. It was found that the Sewing Instructions text and the Pattern Envelope text were similar from the point of view of reference because neither of these text types had any referential pronouns. This is an obvious plus because resolution of anaphoric pronouns is acknowledged as being problematic for MT. The Textbook corpus, on the other hand, displayed all but one type of referential pronoun. On examination of some of the occurrences of pronouns in context in the Textbook corpus, it was found that some did indeed cause ambiguity.

The examination of compounds revealed that the Sewing Instructions had almost double the number of compounds than the other two (31.44% of the total number of words in the corpus in comparison with 18.43% and 15.04% in the Pattern Envelope and Textbook respectively). Moreover, there was a higher degree of repetition of compounds in the Sewing Instructions corpus. When the compounds were analysed for degrees of complexity, the Pattern Envelope and Textbook text types were similar, having approximately 90% "simple" compounds and 10% "complex" each, while the Sewing Instructions differed with 56% "simple", 44% "complex" and 1% "more complex". We conclude, therefore, that the Pattern Envelope and Textbook share

similarities in compounding but the Sewing Instructions differ from the other two text types in this respect.

4.2 Syntactic Ambiguity: Conclusions

The three phenomena analysed under this heading were conjunction, ellipsis and verb forms. The highest number of conjunctions occurred in the Textbook corpus. It also exhibited the widest range of category of conjunction, suggesting that it has a wider semantic range than the other two text types. The degree of coordination in each corpus was similar. However, the same is not true for subordination which occurred mostly in the Textbook text type, followed by the Sewing Instructions. The Pattern Envelope corpus displayed no examples of subordination. Interestingly, although the Pattern Envelope has the lowest number of conjunctions in general, it has the highest number of ambiguities resulting from conjunction. We concluded that in comparison with the number of conjunctions that occur between the three corpora, the number of resulting ambiguities is low and most of them can be disambiguated either by using co-occurrence information or by "controlling" the language to avoid ambiguities.

The next part of the analysis consisted of an examination of ellipsis from a sample of each text type. It was found that the Sewing Instructions had the most examples of ellipsis (32% of total number of words in the sample), followed by the Textbook corpus (14.2%) and then the Pattern Envelope (6.8%). This was surprising because, as already mentioned in the description of each text type, the Pattern Envelope is the most telegraphic of the three text types under examination. We attribute the low occurrence of ellipsis to the fact that few full sentences appear in this text type. Thus, there is little opportunity for ellipsis within the phrases that are used. The high occurrence of ellipsis in the Sewing Instructions corpus is not a good indication for MT because ellipsis is

highly problematic when translating into languages such as French or German where, frequently, the elided article must be restored.

Finally, verb forms used in each text type were compared. Here, the Sewing Instructions and Pattern Envelope texts were similar with the exception that future tense and infinitives are used in the Sewing Instructions whereas they are not used in the Pattern Envelope and, similarly, the Pattern Envelope uses modals but the Sewing Instructions does not. The Textbook exhibited the widest range of verb forms. A striking similarity between the three text types is that the highest occurring verb form is the imperative, accounting for 44% in the Pattern Envelope, 53% in the Textbook and 69% in the Sewing Instructions. This can be attributed to the "instructive" communicative function of all three text types.

4.3 Lexical Ambiguity: Conclusions

Under the heading of lexical ambiguity, the texts were examined for polysemy and grammatical category. It was found that the Textbook had the highest occurrence of multicategorial items (8.13% of the total number of words), followed by the Pattern Envelope (3.22%) and then by the Sewing Instructions (2.15%). In this respect, the Textbook would be the most problematic for MT. However, it was also noted that syntactic information usually provides the necessary information for identification of grammatical category.

Finally, the analysis of polysemy revealed that, once again, the Textbook text had the highest number of polysemes (0.6%), followed by the Pattern Envelope (0.1%) and then the Sewing Instructions (0.09%). As can be deduced from these figures the occurrence of polysemy in these texts in general is very low, not reaching even 1% of the total number of words in any case. Once again, analysis of the syntactic features of each polyseme allowed for disambiguation.

4.4 Conclusions and Recommendations

Although our corpora are not fully representative of each text type, we do feel that they are exemplary and have provided us with an adequate sample of the linguistic characteristics of each text type. On the basis of the results of our analyses, we conclude that it is not correct to speak simply of the linguistic characteristics of "the sublanguage of sewing" because within this sublanguage there are different text types (perhaps more than we have examined) which are characterised by different linguistic features. We have illustrated that our three text types show distinct similarities in only two of the seven linguistic features examined: the Pattern Envelope and Sewing

Instructions are similar because they make no use of referential pronouns and the Pattern Envelope and Textbook display similarities in the number and type of compounds they use. These two categories aside, the texts exhibit distinct linguistic differences although they have similar communicative functions and belong to the same highly restricted subject domian. We would therefore recommend that sublanguage analyses for MT should give careful consideration to the different *types* of text in a sublanguage since it is likely that other sublanguages also have different text types. In the future, it may be possible to sensitise MT grammars to the linguistic (and even extra-linguistic) characteristics of particular text types.

Our description of the macrostructure of each text type, which highlights the fact that both the Pattern Envelope and Textbook text types have descriptive and instructive passages, has proven the claim that within one text type there may be more than one type of text. This leads us on to the second aim of our analysis which was to demonstrate that there is no one rigidly defined text type which can be labelled "instructive". All of our text types have an "instructive" function, yet they differ linguistically. This, we believe, has to do with the extra-linguistic features of each text type which are listed below:

Textbook

- Used in schools
- Expert to Non-expert Communication
- Didactic/Instructive

Pattern Envelope

- Used primarily in shop where fabric and "notions" are bought
- Expert to Expert Communication

- Instructive/Descriptive

Sewing Instructions

- Used in work situation (usually in the home)
- Expert to Expert Communication
- Instructive

We conclude, therefore,that there is no one "instructive" text type but that several texts may be instructive while differing linguistically due to multi-functionality and extra-linguistic considerations such as the knowledge level of the recipient and assumptions made by the author. Our conclusions further substantiate the claims (reported in Chapter 1) made by scholars such as Wilss, for example, that all texts are multi-functional and by others such as Sager, Hatim and Mason and Stolze who claim that the form and function of the text varies according to the knowledge level of the recipient. Furthermore, our conclusions also support claims made by Hatim and Mason and Stolze that there is more than one type of "instructive" text.

Regarding the question of which text type of the three is most suitable for MT, we conclude that no one text type can be identified as being "most suitable" because, while one feature deemed problematic for MT might be absent in one text, that same text may be characterised by another problematic feature. This is the case with the Sewing Instructions which has no anaphoric pronouns, but which has the highest occurrence of ellipsis and compounding. Further, while the Textbook has referential pronouns, it has a lower occurrence of compounding and ellipsis. On the other hand, the Textbook also exhibits the most multi-categorial items and polysemes.[95]

95However, as already pointed, out a sophisticated grammar could recognise these and disambiguate them.

Like the Sewing Instructions, the Pattern Envelope has no referential pronouns. However, it does have the highest number of ambiguities resulting from conjunction. Nevertheless, since the Pattern Envelope text consists of highly telegraphic language with few complete sentences, many tables of figures and a high recurrence of the same syntactic constructions and lexical items, it would be a good candidate for an example-based MT system since it would be neither a difficult nor a time-consuming task to establish TL equivalents for SL structures using a bi-lingual corpus. This is a project which could be an object of investigation in the future.

Developing an MT system for automatically translating the Sewing Instructions text type and the Textbook text type would obviously be a bigger task, but the benefits of developing a working system would be more significant than for the Pattern Envelope text. A bi-lingual corpus would have to be compiled which would be representative of all possible syntactic constructions and lexical classes in the sublanguage.

On a more general level, we have seen that consideration and comparison of text types yields invaluable information about variation between text types in the sublanguage of sewing and we would urge researchers in the domain of machine translation to consider text type, communicative function and extra-linguistic features in the future.

APPENDICES

APPENDIX A
SAMPLE OF SEWING INSTRUCTIONS TEXT

Construction Instructions

VIEW 1

1. NOTE: Determine finished length before sewing

STEP 1: Right Front

Right side uppermost and matching small circles, make tucks at

upper edge of RIGHT front skirt in direction of arrows indicated on pattern.

Tack to hold.

MACHINE NEATEN edge of attached facing

STEP 2 : Left Front

STAY-STITCH upper edge of left front skirt.

Matching stitching lines and circles, stitch dart. Snip to circle as shown.

Press towards centre front.

STEP 3 : Back

STAY-STITCH upper edge of back pieces.

Matching stitching lines and circles, stitch dart.

Snip to circle as shown.

Press towards centre back seam.

Stitch centre back seam. Press.

Right sides together, pin RIGHT front to back at right side seam, matching circles.

Stitch, leaving an opening for sash between circles.

For extra strength, work a few back stitches at circles.

Press.

Right sides together, pin LEFT front to back at left side seam.

Stitch.

APPENDIX B
SAMPLE OF TEXTBOOK TEXT

Slip-tacking or Basting

Slip-basting is a temporary stitch used (i) to hold matching stripes and plaids of seams in position, (ii) on complicated curved seam sections of garments, (iii) when fitting and altering garments on the right side.

Method

1. Mark seam allowance on garment pieces.
2. On upper layer of garment fabric turn seam allowance to wrong side.
3. Place fitting line of upper layer to fitting line of matching garment piece. Pin to hold in position.
4. With right sides facing, working from right to left, slip the needle between folds of fabric to conceal thread and bring out at upper folded edge.
5. Repeat this process. Fasten off threads with a backstitch.

Machine Tacking

This is suitable for firm, closely woven fabrics that will not
gather or mark. Fine fabrics tend to form gathers when machine tacked.

Method

1. Pin seams together matching notches, fitting lines and placing pins at right angles to raw edge of seam.
2. Set machine at its longest straight stitch. The upper tension may be loosened slightly to ensure that the machining is easy to remove.
3. Using a contrasting colour of thread, machine along the seam 1 mm inside the seamline.
4. When permanent stitching has been completed unpick the machine tacking. cut thread end 30 mm away from loop. The thread ends should always be longer than loop. Repeat this process at all balance marks. Remove pins.
5. Remove each pattern piece by pulling uncut loops through, leaving a small tear in the pattern, or cutting the loops, removing pattern and leaving pattern without any tears. Roll the fabric away from the fabric gently.
6. Separate fabric layers by pulling apart carefully and cutting the threads between the layers. Each garment piece is marked with tufts of thread on balance marks.
7. Fold each garment piece to prevent tailor tacking coming out.

APPENDIX C
SAMPLE OF PATTERN ENVELOPE TEXT

Dress with gathers from dropped waistline has front button fastening. View 1 with extended shoulders has shoulder yoke, side seam pockets and three-quarter length sleeves with turn back cuff. Worn with purchased belt. View 2 is sleeveless with wide neckline and patch pockets from side seams.

6512 FIVE SIZES IN ONE

SIZE BUST

VIEW 1 (Measurements)

90cm**

115cm** (Measurements)

150cm*/**

VIEW 2

90cm**

115cm*/* (Measurements)

150cm*/**

INTERFACING: BOTH VIEWS

80cm

Allow extra fabric for matching plaids or stripes.

* with nap, shading, pile or with a one-way design.

** Without nap, shading, pile or with a two-way design.

REQUIREMENTS: BOTH VIEWS, five 13mm (½") buttons. VIEW 1, purchased belt.

SUGGESTED FABRICS; Cotton, Linen, Seersucker, Broderie Anglaise, Viscose, Fine Needlecord, Soft Denim, Synthetics.

STANDARD BODY MEASUREMENTS:

SIZE:

BUST: (Measurements)

WAIST:

HIP:

Nape to Waist:

Nape to finished length both views:

Finished width round hem:

APPENDIX D
COMPOUNDS IN SEWING INSTRUCTIONS

Compound	*No. of Occurrences*
armhole edge	02
armhole edges	05
armhole seam	01
armhole seams	01
armhole seam allowance	01
attached facing	03
back bodice pieces	01
back bodice panel seams	02
back bolero	02
back facing	02
back facing pieces	03
back lining	07
back lining section	01
back lining section(s)	01
back neck edge	01
back neck edges	01
back neck facing pieces	03
back neck seam	01
back opening	01
back opening edges	02
back pieces	02
back seam allowance	02
back sections	03
back shoulder edges	01

back skirt pieces	01
back stitches	06
back top pieces	01
basted edges	01
bodice seams	01
bodice seam allowance	01
bodice shoulder	01
bolero side seams	01
button holes	01
center back	06
centre back	03
centre back edges	02
centre back opening edges	01
centre back seam	05
centre front	02
centre front circle	02
centre front circles	05
centre front lines	01
collar edges	02
construction instructions	04
continuous operation	02
dart edges	01
double notches	01
extra strength	08
facing centre back seam allowance	02
facing pieces	07
facing sections	01
facing shoulder seams	02
facing side seams	01

finished edges	02
finished length	02
flap sections	01
folded edge	03
fold line	04
front bodice	02
front bodice panel seams	02
front bolero	02
front bolero piece	01
front bolero pieces	01
front edge	03
front facing	04
front facings	02
front facing piece	01
front facing pieces	01
front interfacing	03
front lining	02
front lining sections	01
front neck edge	03
front neck facing piece	01
front opening	03
front opening edge	03
front opening edges	01
front panel seams	01
front sections	02
front skirt	01
front skirt side seam allowance	01
garment neck edge	01
hem allowance	04

hem edges	01
hem line	01
inner corners	01
inner edge	02
interfacing seam allowances	01
lapel edges	02
large centre back circle	01
large centre back circles	01
large centre front circles	05
large circles	01
large dart	01
larger sizes	01
left back	02
left back dress	01
left back edge	01
left back waistband	02
left facing	01
left front	19
left front edge	02
LEFT front facing	01
left front opening edge	02
left front skirt	02
left side seam	04
LEFT tie	03
lining back panel	01
lining side seam	01
lining side seams	01
long edge	02
long edges	01

long machine stitches	01
long running stitches	01
long sleeve	01
lower edge	20
lower edges	03
lower front opening	02
lower pocket	03
lower raw edges	01
Lower Welt 3	02
Lower Welt 4	01
matching fabric section	03
narrow hem	06
neatened edge	03
neck edge	16
neck edges	04
neck facing	01
neck seams	01
neck seam	01
notched edges	01
notched ends	02
one pair	01
opening edge	04
opening edges	06
outer circles	01
outer edge	01
outer edges	01
panel seam	01
panel seams	11
placement line	01

pocket bag	02
pocket bags	03
pocket edges	01
pocket ends	01
pocket facing	03
pocket opening	02
pointed end	01
pressed edge	03
pressed edges	03
previous stitching	04
raw edge	05
raw edges	11
right back edge	02
right back skirt	01
right back stitching	01
RIGHT back waistband	01
right facing	03
Right Front	32
right front attached facing	02
RIGHT front facing	02
right front opening edge	01
RIGHT front skirt	02
right front welt	01
right side	18
right sides	21
right side seam	03
right side seam opening	01
RIGHT TIE	03
right tie end	01

roll line	05
seam allowance	19
seam line	04
seam tape	02
second row	01
self facing	04
shoulder edge	01
shoulder pad	01
shoulder pads	01
shoulder pad placement	01
shoulder seam	09
shoulder seams	20
side back bolero pieces	01
side front bolero	01
side front lining	01
side front lining pieces	01
side seam	07
side seams	20
side seam edges	02
sleeve facing seam	01
sleeve head	03
sleeve interfacing	01
sleeve seam	04
sleeve seam allowance	01
sleeve seams	01
small circles	08
small dart(s)	01
stable fabrics	02
stitching line	06

stitching lines	11
stitching lines inner corners	02
thread shank	01
thread shanks	01
thread ends	02
tie end	11
tie end seam allowance	02
tissue pattern	02
triangular ends	05
underarm curve	01
underarm seam	01
underarm seams	02
under collar sections	01
under sleeve lining	01
unnotched edges	01
unstable fabrics	02
upper collar	05
upper collar seam allowance	02
upper edge	31
upper edges	01
upper pocket	01
upper raw edges	01
upper sleeve	01
upper sleeve lining	01
upper welt	01
vest side seams	01
View 1 step 4	02
View 1 steps 3,4,5 and 6*	01
View 1 step 8*	01

View 2 steps 1 to 6*	01
View 2 & 3	01
waistband extension	01
wearing ease	01
welt seam	01
welt section	01
wrong side	11
wrong sides	03
zipper foot	02
zip tape	02
zip teeth	02

APPENDIX E
COMPOUNDS IN TEXTBOOK CORPUS

Compound	*No. of Occurrences*
Assissi cross stitch	01
back stitches	01
balance marks	07
bias binding	04
blanket stitch	06
blanket stitching	03
blind hemming	01
buttonhole stitch	01
centre back	03
centre back line	01
centre front	01
centre line	02
centre point	01
centreback lines	01
chain stitch	02
chain stitches	01
chalk pencil	01
Chanel seam	02
crease line	01
crossed seams	01
crossed stitches	02
curved seams	03
curved seam sections	01
daisy stitch	01

decorative finish	02
design line	04
design outline	01
diagonal stitch	03
diagonal stitches	02
diagonal tacking	01
double backstitch	09
double feather stitch	01
double pointed dart	01
double welt seam	01
easing stitch	01
edge-matching	01
embroidery thread	02
embroidery threads	03
embroidery stitches	04
fabric fold	02
fabric edges	01
fashion feature	01
fabric layers	01
feather stitch	01
felled seams	01
first finger	02
fitting line	22
flat finish	01
flat seam	03
fly stitch	02
flat stitches	01
folded edge	12
French knot	01

machine gathering	01
machine stitching	02
machine tacking	02
machine felled seam	01
matching stripes	02
middle finger	01
outward corner	01
outward curves	01
overlay pieces	01
Paris binding	01
pattern marking	01
pattern markings	01
pattern piece	03
pinking shears	01
plain seam	01
pocket seams	01
raw edge	15
right angles	06
right hand	01
right sides	09
run and fell seam	03
running stitches	09
saddle stitch	02
satin stitch	01
satin stitches	02
seam allowance	30
seam fitting line	02
second finger	01
self-finished seams	02

APPENDIX F
COMPOUNDS: PATTERN ENVELOPE

back tucks	01
bias collar	01
bias binding	02
body measurements	03
Broderie Anglaise	02
button cuffs	01
button trim	01
contrast collar	02
contrast scarf	02
Cotton Sateen	01
Cotton Mixtures	04
Crêpe de Chine	02
dropped waistline	01
edging lace	01
extended shoulders	03
Fine Needlecord	01
front button fastening	01
hemline vent	01
hook and eye	02
hooks and eyes	01
instruction sheet	02
Lightweight Wool	01
Lightweight Silk	01
loose-fitting	02
lower edge	02
Lt. Wt. Melton	01

Lt. Wt. Woolens	02
Lt. Wt. Double Knits	02
matching plaids	06
mid-knee	02
mock vent	01
nap yardages/layouts	02
one-way design	06
one-way design fabrics	02
patch pockets	01
pattern measurements	04
pleated right front	01
Polyester Cotton Prints	01
raised waist	02
semi-fitted	01
semi-fitting	01
shoulder yoke	01
shoulder pads	07
side button front	01
side panels	01
side seam pockets	01
side seams	02
Silk-Like Linen	02
Soft Denim	01
stitched facings	01
three-quarter length	01
tie end	02
Tissue Faille	01
turn back cuff	01
Twilled Cotton	01

APPENDIX G
ELLIPSIS - SEWING INSTRUCTIONS

Occurrences of Ellipsis in Sewing Instructions sample:

Determine Ø (the) finished length before sewing Ø (the garment).

Ø (With the) Right side uppermost and matching Ø (the) small circles, make tucks at Ø (the) upper edge of Ø (the) Right front skirt in Ø (the) direction of Ø(the) arrows Ø(as is) indicated on Ø(the) pattern.

Tack Ø(the material) to hold Ø(it).

MACHINE NEATEN Ø(the) edge of Ø(the) attached facing.

STAY-STITCH Ø(the) upper edge of Ø(the) left front skirt.

Matching Ø(the) stitching lines and Ø(the) circles, stitch Ø(the) dart.

Snip Ø(the material) to Ø(the) circles as Ø(is) shown.

Press Ø(the material) towards Ø(the) centre back seam.

Stitch Ø(the) centre back seam.

Press Ø(the material).

Ø(With the) Right sides together, pin Ø(the) RIGHT front to Ø(the) back at Ø(the) right side seam, matching Ø(the) circles.

Stitch Ø(the material), leaving an opening for Ø(the) sash between Ø(the) circles.

For extra strength, work a few back stitches at Ø(the) circles.

Press Ø(the material).

Ø(With the) Right sides together, pin Ø(the) LEFT front to Ø(the) back at Ø(the) left side seam.

Stitch Ø(the material).

Fold Ø(the) RIGHT tie lengthways with Ø(the) right sides together.

Stitch Ø(the material) as Ø(is) shown.

Trim Ø(the) corners and Ø(the) seams.

Turn Ø(the material) to Ø(the) right side.

Press Ø(the material).

Ø(With the) right sides uppermost, pin Ø(the) left tie to Ø(the) left front, matching Ø(the) small and Ø(the) large circles.

Tack Ø(the material).

Fold Ø(the) left tie in half lengthways, with Ø(the) right sides together.

Trim Ø(the) corners and Ø(the) seams.

Turn to Ø(the) right side.

Press Ø(the material).

Ø(With the) Right sides uppermost, pin Ø(the) LEFT TIE to LEFT front, between Ø(the) small circles as Ø(is) shown.

Tack Ø(the material.)

Apply Ø(the) interfacing to Ø(the) facing pieces.

Pin Ø(the) back facing to Ø(the) RIGHT front facing at Ø(the) right side seam, matching Ø(the) small circles.

Stitch Ø(the material), leaving an opening between Ø(the) circles.

For extra strength, work a few back stitches at Ø(the) circles.
Pin Ø(the) left front facing to Ø(the) back facing at Ø(the) LEFT side seam, matching Ø(the) double notches.

Stitch Ø(the material).

Press Ø(the material).

MACHINE NEATEN Ø(the) lower edge of Ø(the) facing.

Ø(With the) right sides together, pin Ø(the) facing to Ø(the) skirt, matching Ø(the) side seams and Ø(the) large and Ø(the) small circles.

Stitch Ø(the material), stitching through Ø(the) large and Ø(the) small circles at Ø(the) right front and along Ø(the) foldline as Ø(is) shown.

Trim Ø(the) corners.

Layer Ø(the) seams.

Turn Ø(the) facing to Ø(the) inside, turning Ø(the) right front attached facing to Ø(the) inside along Ø(the) fold line.

Press Ø(the material).

Tack Ø(the) facing to Ø(the) side seams, Ø(the) darts and Ø(the) right facing.

Catch-stitch around Ø(the) right side seam opening.

Press under 1.5 cm Ø(of material) on Ø(the) left front edge.

Tuck under Ø(the) raw edge to meet Ø(the) crease.

Stitch Ø(the material) in position.

Press Ø(the material).

Ø(With the) Right side uppermost, top-stitch 6mm from Ø(the) upper edge as Ø(is) shown.

Turn Ø(the) right front attached facing onto Ø(the) right side.

Stitch across Ø(the) facing along Ø(the) hem line.

Trim Ø(the) facing close to Ø(the) stitching. Trim Ø(the) skirt to within 1.5 cm of Ø(the) inner edge of Ø(the) facing.

Turn Ø(the) facing to Ø(the) inside.

Press Ø(the material).

Machine neaten Ø(the) lower edge of Ø(the) skirt.

Turn up Ø(the) hem allowance and tack Ø(the material) to hold Ø(it) in position.

Turn back Ø(the) neatened edge and loosely catchstitch Ø(the) hem to Ø(the) skirt.

Avoid Ø(the) stitches showing on Ø(the) right side.

Press Ø(the material).

Determine Ø(the) finished length before sewing Ø(the garment).

With Ø(the) right side uppermost and matching Ø(the) circles, make tucks at Ø(the) upper edge of Ø(the) right front skirt in Ø(the) direction of Ø(the) arrows Ø(as is) indicated on Ø(the) pattern.

APPENDIX H
ELLIPSIS: TEXTBOOK:

Occurrences of Ellipsis in Textbook Sample:

However, fine hand-sewing is necessary for many parts of a garment e.g. hemming Ø (the) waistband to (a) row of Ø (the) machine, top-sewing Ø (the) ends of Ø (the) waistband, slip-hemming Ø (the) hems of Ø (the) garments etc.

Select the right type of needle and Ø (the right type of) thread. The colour and Ø (the) fibre in the thread should match those Ø(colours and fibres) of the garment fabric.

5. Place a thimble on the middle finger of the right hand. It should be used when Ø (you are) hand sewing. The blunt end of the thimble helps when Ø (you are) pushing the needle through the fabric.

Some people find it easier to sew from Ø(the) left to Ø(the) right especially people who are left-handed.

6. Always pin Ø (the material) before tacking Ø(the) garment pieces, Ø(the) hems etc. together.

7. Begin temporary stitches with a knot to secure the thread end.

8. Begin permanent stitches with one or two back stitches concealing the thread end between Ø(the) fabric fold.

When the permanent stitches have been completed, Ø(the) temporary stitches should be easy to remove.

10. Work Ø(the) stitches evenly and accurately. Do not pull Ø(the) stitches too tightly.

Work Ø(the) backstitching to fasten off Ø(the) thread securely and conceal Ø(the) thread end in Ø(the) fabric fold (Check Ø(the) instructions on Ø(the) stitches for exceptions).

12. Remove Ø(the) tacking and press Ø(the) completed permanent stitches.

(i) To hold matching stripes, plaids, balance marks, notches etc. of two garment pieces together for Ø(the) first fitting and Ø(for)machine Ø(stitching) or hand stitching.

(ii) To attach Ø(the) interfacing to Ø(the) garment piece and hold it in position during machining.

(iii) Used near Ø(the) fitting line, it is a good guide when stitching.

Note: Before tacking, pin Ø(the) garment pieces together matching Ø(the) notches, Ø(and the) fitting line etc. Use a thread of Ø(a) contrasting colour, making sure that it will not mark the fabric, e.g. never use black on white.

Ø(It/this is used for)Marking Ø(the) centrefront and Ø(the) centreback lines of Ø(a/the)garment.

Ø(It/this is used for) Holding Ø(the) darts, Ø(the) seams and Ø(the) hems in position during fitting and machining.

Ø(It/this is used for)Attaching Ø(the) interfacings and Ø(the) linings.

Ø(It/this is used for) Keeping Ø(the) lower edge of Ø(the) hem in position until Ø(garment/garment name is) completed.

1. Match Ø(the) notches and Ø(the) edges of Ø(the) garment. Place Ø(the) pins at right angles to Ø(the) fabric edges.

2. Begin by using a knot or Ø(a) double backstitch.

3. Always work from Ø(the) right to Ø(the) left. Make Ø(the) stitches and Ø(the) spaces equal in size, (10 mm to 12 mm approx.) for even tacking. For uneven tacking work long stitches (15 mm approx.) with small spaces between them (10 mm approx.) on Ø(the) side facing you.

4. Fasten off Ø(the) thread securely with a double backstitch.

This type of tacking is suitable for keeping a few layers of fabric in position until Ø(the) fitting and Ø(the) machining have been completed, e.g. on collars, facings, interfacings, linings and fine fabrics.

2. Always work from Ø(the) right to Ø(the) left.

3. Fasten off Ø(the) thread securely with a double backstitch.

Slip-basting is a temporary stitch used (i) to hold matching stripes and plaids of Ø(the) seams in position,

1. Mark Ø(the) seam allowance on Ø(the) garment pieces

2. On Ø(the) upper layer of Ø(the) garment fabric turn Ø(the) seam allowance to (the) Øwrong side.

3. Place Ø(the) fitting line of Ø(the) upper layer to Ø(the) fitting line of Ø(the) matching garment piece. Pin Ø (it/material) to hold in position.

4. With Ø(the) right sides facing, working from Ø(the) right to Ø(the) left, slip the needle between Ø(the) folds of Ø(the) fabric to conceal Ø(the) thread and bring Ø(it) out at Ø(the) upper folded edge.

5. Repeat this process. Fasten off Ø(the) threads with a backstitch.

This is suitable for firm, closely woven fabrics that will not gather or mark. Fine fabrics tend to form gathers when Ø(they are) machine tacked.

APPENDIX I
ELLIPSIS: PATTERN ENVELOPE

Occurrences of Ellipsis in Pattern Envelope Sample:

Ø(This is a) Dress with gathers from Ø(the) dropped waistline Ø(which) has front button fastening.

View 1 with Ø(the)extended shoulders has Ø(a)shoulder yoke, side seam pockets and three-quarter length sleeves with Ø(a)turn back cuff.

Ø(It is) Worn with Ø(a) purchased belt. View 2 is sleeveless with Ø(a)wide neckline and patch pockets from Ø(the) side seams.

* with Ø(a) nap, shading, pile or with a one-way design.
** Without Ø(a) nap, shading, pile or with a two-way design.

REQUIREMENTS: BOTH VIEWS, five 13mm (½") buttons. VIEW 1, Ø(a) purchased belt.

Ø(From the) Nape to Ø(the) Waist

Ø(from the) Nape to Ø(the) finished length Ø(of) both views

Ø(From the) Nape to Ø(the) finished length of Both Views

1 CONTRAST COLLAR and VIEW 2 Ø(has an) OPTIONAL CONTRAST SCARF Ø(which is) suitable for soft draping fabrics only such as Crêpe de Chine, Voile, Organza, Lightweight Silk, Washed Silk and Chiffon.

Ø(From the) Nape to Ø(the) finished length of view 1& 2

Ø(From the) Nape to Ø(the) finished length of View 1 & 2

Ø(From the) Nape to Ø(the) finished length of View 3

Ø(from the) Nape to Ø(the) finished length

Ø(The) Finished length of View 1 from Ø(the) waistline

Ø(The) Finished length of View 2 including Ø(the)waistband

Ø(The) Finished length of View 3 excluding Ø(the) waistband

Ø(from the) Nape to Ø(the) approximate finished length of Both View Dresses

Ø(From the) Nape to Ø(the) finished length of Ø(the) Bolero

Ø(A) MISSES' JACKET, SKIRT & BLOUSE Ø (which is) Loose-fitting, lined, above Ø(the)hip jacket has extended shoulders, shoulder pads, welt pockets with flaps, side panels (no side seams) and long two-piece sleeves with Ø(a) mock vent and button trim.

Ø(The) Semi-fitted, straight skirt, Ø(which falls) below mid-knee, has Ø(a) raised waist and Ø(a) back zipper and Ø(a) hemline vent.

Ø(The) Very loose-fitting blouse has extended shoulders, shoulder pads, Ø(a) bias collar, front and back tucks and long sleeves with button cuffs.
Ø(It has a) Narrow hem.

Ø(It/the garment is) Unsuitable for obvious diagonals, plaids or stripes.

WIDTHS/LENGTHS: Ø(the) Width at Ø(the) Lower Edge: Jacket 38 to 52" (96 to 132cm) Skirt 36 to 49" (91 to 124 cm), Blouse 38 to 51" (96 to 130cm).

Ø(the) Back length from Ø(the) base of Ø(the) neck: Jacket 22¼ to 24¼" (56.5 to 61.5cm)

Ø(The) Back Length from Ø(the) Waist: Skirt 24Ø"(63cm)

See Ø(the) Instruction Sheet for 35" (90cm) Yardage.

Ø(The) MISSES' SKIRT Ø(is a) Semi-fitting, wrap skirt, Ø(which falls) above mid-knee, Ø(and) has Ø(a)raised waist with stitched facings.

A: Ø(with) side button front, and welt pocket. B: pleated, rightside with Ø(a)drape (wrong side may show)

C:Ø(a)pleated right front extends into Ø(the) tie end, and Ø(the)left tie end (wrong side may show).

B:and C: Ø(have a) narrow hem. Purchased top.

Skirt C: Ø(This is) Unsuitable for obvious diagonals, plaids or stripes.

See Ø(the) instruction Sheet for 35£ (90cm) yardage.

APPENDIX J
MULTI-CATEGORIAL ITEMS: SEWING INSTRUCTIONS

Lexical Item	*Grammatical Categories*[96]
back	N / N (mod) / ADV
collar	N / N (mod)
ease	N / V
flap	N / N (mod)
fold	N / N (mod) / V
hand	N / N (mod / V
hem	N / N (mod)
layer	N / V
lining	N / N (mod)
machine	N (mod) / V
matching	V / V (mod)
open	ADJ / V / ADV
opening	N / V (mod)
panel	N / V (mod)
place	V / ADV
pleat	N / N (mod)
seam	N / N (mod)
skirt	N / N (mod)
sleeve	N / N (mod)
thread	V / N (mod)
top	N / N (mod)

[96] N = noun/ N (Mod) = modifying noun/ V = verb/ V (Mod) = verb modifying a noun/ ADV = adverb/ ADJ = adjective

waistband	N / N (mod)
welt	N / N (mod)
zip	N / N (mod)

APPENDIX K
MULTI-CATEGORIAL ITEMS:TEXTBOOK CORPUS

Lexical Item	*Grammatical Category*
armhole	N/N (mod)
back	N/N (mod)/Adv
beginning	N/V
binding	N/V
centre	N/Adj
completed	V/V (mod)
crease	N/N (mod)
cross	N/N (mod)/V
cut	N/V
cutting	N/V
deep	Adv/Adj
design	N/N (mod)
easing	V/V (mod)
embroidery	N/N (mod)
even	Adj/Adv
fabric	N/N (mod)
finish	V/N
finished	V/V (mod)
fitting	N/N (mod)/V
flat	Adv/Adj
fold	N/V
form	V/N
front	N/N (mod)

garment	N/N (mod)
hand	N/N (mod)/Adv
hem	N/N (mod)/V
hemming	V/N
layer	N/V
light	N/Adj
loop	N/N (mod)
machine	N/N (mod)/V
marking	V/N
matching	V/V (mod)
outline	V/N
outlines	V/N
outside	Adj/Adv
overcasting	N/V
overlay	N/N (mod)
pattern	N/N (mod)
pin	N/V
pinking	N/N(mod)
place	V/Adv
point	N/N (mod)
points	V/N
repeat	V/V (mod)
right	N/N (mod)
run	V/V (mod)
running	N/V (mod)/V
seam	N/N (mod)/V
sewing	N/N (mod)/V (mod)
slip-hemming	N/V
space	N/V

starting	V/V (mod)
stitch	N/V
stitching	N/V/V (mod)
tacking	N/V/V (mod)
thread	N/N (mod)
top-sewing	N/V
trimmed	V/V (mod)
underlay	N/N (mod)
waist	N/N (mod)
work	N/V
zig-zag	V/Adj

APPENDIX L
MULTI-CATEGORIAL ITEMS:PATTERN ENVELOPE CORPUS

Lexical Item	*Grammatical Category*
back	N/N (mod)
button	N/N (mod)
cotton	N/N (mod)
design	N/N (mod)
finished	V/V (mod)
front	N/N (mod)
hip	N/N (mod)
length	N/N (mod)
nap	N/N (mod)
side	N/N (mod)
wool	N/N (mod)

BIBLIOGRAPHY

Aarts, J. & W. Meijs (eds.), *Corpus Compilation and the Automatic Analysis of English*, (Amsterdam: Rodopi, 1991)

Adler, S., "The Birth of a Standard", in *Journal of the American Society for Information Science*, (43(8), 1992), pp. 556-558

Atkins, S., J. Clear, N. Ostler, "Corpus Design Criteria", in *Literary and Linguistic Computing*, (Vol. 7, no. 2, 1992), pp.1-16

Arntz, R. & G. Thome (eds.), *Übersetzungswissenschaft: Ergebnisse und Perspektive (Festschrift für Wolfram Wilss)*, (Tübingen: Gunter Narr Verlag, 1990)

Arntz, R. (ed.), *Textlinguistik und Fachsprache*: Akten des Internationalen übersetzungswissentschaftlichen AILA-Symposiums Hildesheim, 13 - 16 April 1987, (Zürich, New York: Georg Olms, 1988)

Baumann, K.D., "Die Makrostruktur von Fachtexten", in *Special Language - Fachsprache*, (Heft 1-2, 1987), pp.2-18

Beaman, K., "Co-ordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse", in *Series in Advances in Discourse Processes*, D. Tannen (ed.), (New Jersey: Ablex Publishing Co., 1984), pp. 45-80

Beaugrande, R de, W. Dressler, *An Introduction to Text Linguistics*, (New York: Longman, 1981)

Beaugrande, R de, "Text Linguistics and New Applications", in *Annual Review of Applied Linguistics*, (vol. 11, 1990), pp. 17-41

Beedham, C., M. Bloor, "English for Computer Science and the Formal Realization of Communicative Functions", in *Special Language-Fachsprache*, (Vol.11, Issue 1-2, 1989), pp. 13-24

Biber, D., "Investigating Macroscopic Textual Variation through Multi-feature/Multi-dimensional Analyses", in *Linguistics*, (Vol. 23, 1985), pp. 337-360

Biber, D., "A Typology of English Texts", in *Linguistics 27*, 1989, pp. 3-43

Biber, D., *Variation Across Speech and Writing*, (Cambridge & New York: Cambridge University Press, 1989)

Bonzi, S., "Syntactic Patterns in Scientific Sublanguages - A Study of Four Disciplines", in *Journal of the American Society for Information Science* 5, (Vol.41 (2), 1990), pp. 121-131

Brinker, K., *Linguistische Textanalyse*, (Berlin: Erich Schmidt Verlag, 1985)

Brinker, K., "Bedingungen der Textualität", in *Der Deutschunterricht*, (Vol. 40, Pt. 3, 1988), pp. 6-19

Brown, G., G. Yule, *Discourse Analysis*, (Cambridge: Cambridge University Press, 1983)

Brown, P. et al., *A Statistical Approach to MT*, Research Report from IBM Research Division, (New York, 1989)

Bryan, M., *SGML: An author's guide to the standard generalized markup language*, (Wokingham & Massachusetts: Addison-Wesley, 1988)

Buchmann, B., S. Warwick, P. Shann, "Design of a Machine Translation System for a Sublanguage", in *Proceedings of Coling*, (2-6 July, 1984), (New Jersey: Association for Computional Linguistics, 1984), pp. 334-337

Bungarten, T., "Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft", in *Empirische Textwissenschaft*, H.Bergenholtz & B.Shaeder (eds.), (Königstein/Ts. Scriptor, 1979), pp.28-51

Butler, C.S. (ed.), *Computers and Written Texts*, (Oxford (U.K.) & Cambridge (U.S.A.): Blackwell, 1992)

Candel, D., "Ambiguité d'Origine Polysémique dans une Langue de Spécialité", in *Cahiers de Lexicologie*, (Vol. 2, 1984), pp.21-32

Chalker, S., *Current English Grammer*, (London: Macmillan, 1984)

Chandioux, J., M. Guéraud, "Météo: un système à l'épreuve du temps", in *META*, (Vol. 26 (1), 1981), pp.18-22

Crocker, M., "A Principle-Based System for Syntactic Analysis", in *Canadian Journal of Linguistics*, (1991), pp.1-26

Crystal, D., *A Dictionary of Linguistics and Phonetics*, (Oxford, New York: Basil Blackwell, 1985)

Efthinionm E. et al., *Final Report on Anaphora*, Eurotra Internal Report, (1989)

Garside, R., G. Leech & G. Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*, (London, New York: Longman, 1987)

Gerisch, P., "Anmerkung zum Passivgebrauch in Fachsprachen", in *Special Language - Fachsprache*, (Heft 3-4, 1986), pp. 169-171

Gläser, R., "The Problem of Style Classification in LSP", in J. Hoedt et al (eds.), *Proceedings of the 3rd European Symposium on LSP: "Pragmatics and LSP"*, (Copenhagen, August, 1981), (Copenhagen: LSP Centre, UNESCO ALSED LSP Network and Newsletter, Copenhagen School of Economics, 1982), pp. 69-82

Gläser, R., "The Concept of LSP Rhetoric in the Framework of Modern Text Linguistics", in *Philogoica Pragensia*, (Vol. 3, 1987), pp.113-119

Grimes, J., "Reference Spaces in Text", in *Nobel Symposium*, (Cornell University and Summer Institute of Linguistics, 1982), pp. 381-414

Grishman, R. et al., "Discovery Procedures for Sublanguage Selectional Patterns - Initial Experiments", *Computational Linguistics* (Vol.12, No.3, 1986)

Grishman, R., and R. Kittredge (eds.) *Analyzing Language in Restricted Domains: Sublanguage Description and Processing,* (Hillsdale, New Jersey and London, Lawrence Erlbaum, 1986)

Halliday, M. & R. Hasan, *Cohesion in English*, (London, New York: Longman, 1976)

Hanks, P. et al. (eds.), *Collins Dictionary of the English Language,* (2nd edition), (Glasgow & London: William Collins & Sons, 1986)

Harris, Z. et al., *The Form of Information in Science: Analysis of an Immunology Sublanguage*, (Dordrecht, Boston, London: Kluwer, 1989)

Harris, Z., *Mathematical Structures of Language*, (New York: Wiley & Sons, 1968)

Hartmann, R.R.K., *Contrastive Textology*, (Heidelberg: Julius Groos Verlag, 1980)

Hauenschild, C., "Textlinguistische Probleme der maschinelle Übersetzung", in *LSP Newsletter*, Vol. 10 2(25), pp. 11-25

Hauenschild, C., "Discourse Structure - Some Implications for Machine Translation", in D.Maxwell et al (eds.), *New Directions in Machine Translation*, Conference Proceedings, Budapest 18-19 August, 1988, (Holland & Providence: Dordrecht), pp. 145-155.

Heinemann, W., "Textlinguistik heute - Entwicklung, Probleme, Aufgaben", in *Wissenschaftliche Zeitschrift der Karl Marx Universität, Leipzig*, (Gesellschafts- / Sprachwissenschaftliche Reihe, Vol. 31, Pt. 3, 1982), pp.210-221

Hirschmann, L., N. Sager, "Automatic Information Formatting of a Medical Sublanguage", in *Sublanguage: Studies of Language in Restricted Semantic Domains*, R. Kittredge & J. Lehrberger (eds.), (Berlin & New York: de Gruyter,1982), pp.27-80

Hirst, G., *Semantic Interpretation and the Resolution of Ambiguity,* (Cambridge: Cambridge University Press, 1987)

Hoffman, L., "Seven Roads to LSP", in *Special Language-Fachsprache* (vols. 1-2, 1984), pp.28-37

Hoffman, L., *Kommunikationsmittel Fachsprache: eine Einführung,* (Tübingen: G. Narr, 1985)

Hoffman, L., "On the Place of LSP Research in Applied Linguistics", in *Special Language-Fachsprache* (vols. 1-2, 1985), pp.2-11

Hoffmann, L., "Ein textlinguistischer Ansatz in der Fachsprachenforschung", in M. Sprissler (ed.), *Standpunkte der Fachsprachenforschung*, (Tübingen: Gunther Narr Verlag, 1987), pp.91-105.

Hoffmann, L., "Makrostruktur und Kohärenz als Fachtextsortenmerkmale", in *Wissenschaftliche Zeitschrift der Karl-Mark Universität Leipzig*, (Gesellsch.-wiss. Reihe, no. 37, 1988), pp. 552-565

Huizhong, Y., "A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts", in *Literary and Linguistic Computing*, (Vol.1, No.2, 1986), pp.93-103

Hutchins, W. J., *Machine Translation: Past, Present, Future,* (Chichester [West Sussex] & New York: Halstead Press, 1986)

Hutchins, W. J., H.L. Somers, *An Introduction to MT*, (London: Academic Press, 1992)

Isabelle, P. & L. Bourbeau, "TAUM-AVIATION: Its Technical Features and some Experimental Results", in *Computational Linguistics*, (Vol. 11, 1985), pp.18-27.

Jones, L. B., "Pragmatic Aspects of English Text Structure", *Summer Institute of Linguistics*, (University of Texas, Arlington, 1983)

Kalverkamper, H., "Fachsprachen und Textsorten", in J. Hoedt et al (eds.), *Proceedings of the 3rd European Symposium on LSP, "Pragmatics and LSP"*, (Copenhagen, August, 1981), (Copenhagen: LSP Centre, UNESCO ALSED LSP Network and Newsletter, Copenhagen School of Economics, 1982), pp. 106-154

King, M. (ed.), *Parsing Natural Language*, (London, New York: Academic Press, 1983)

King, M. et al., "Machine Translation Today: the State of the Art", in *Proceedings of the Third Lugano Tutorial,* Lugano, Switzerland 2-7 April, 1984: (Edinburgh: Edinburgh University Press: 1987)

Kittredge, R. & J. Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains,* (Berlin and New York: Walter de Gruyter, 1982)

Kittredge, R., "Variation and Homogeneity of Sublanguages", in R. Kittredge & J. Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains* (de Gruyter,1982), pp. 107-137

Kittredge, R. et al., "Towards a Computable Model of Meaning Text Relations within a Natural Sublanguage", *Proceedings of the 8th International Joint Conference on AI*, (Los Altos, California: William Kaufmann, 1983), pp.657-659

Kosaka, M., V. Teller, R. Grishman, "A Sublanguage Approach to Japanese-English Machine Translation", in *New Directions in Machine Translation*, Conference Proceedings, Budapest 18 - 19 August, 1988, (Dordrecht: Holland, Providence: RI-USA, Foris Publications), pp.109-121

Kussmaul, P., "Instruktionen in deutschen und englischen Bedienungsanleitungen", in Arntz, R., G. Thome (eds.), *Übersetzungswissenschaft-Ergebnisse und Perspektiven: Festschrift für Wolfram Wilss zum 65. Geburtstag*, (Tübingen: G. Narr, 1990), pp.369-379

Langendoen, T. D., "The Grammatical Analysis of Texts", in *Nobel Symposium*, (Cornell University and Summer Institute of Linguistics, 1982), pp. 161-185

Laurén, C. et al., "Potential Fields of LSP Research: A proposal for corpus selection and methods", in *Multilingua*, (4-4, Mouton, 1985), pp.227-230

Laurén, C. & M. Nordmann (eds.), *Special Languages: From Humans Thinking to Thinking Machines* (Papers presented at the 6th European Symposium on LSP, University of Vaasa, Aug. 3rd-7th, 1987), (Clevedon & Philadelphia, Multilingual Matters, 1989)

Laurén, C. & M. Nordmann, "Corpus Selection in LSP Research", in H. Schröder (ed.), *Subject-oriented Texts - Languages for Special Purposes and Text Theory*, (Berlin: Walter de Gruyter, 1991), pp. 218-230

Leech, G., Garside, R. Atwell, E., "Recent Developments in the Use of Computer Corpora in English Language Research", in *Transactions of the Philological Society*, (1983), pp.23-40

Leech, G., *Semantics*, (Penguin, 1974)

Lehrberger, J. & L. Bourbeau, *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation,* (Amsterdam, Philadelphia: John Benjamins Publishing Co., 1988)

Lehrberger, J., "Sublanguage Analysis", in R. Grishman & R. Kittredge (eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, (Hillsdale, New Jersey & London: Lawrence Erlbaum, 1986), pp.19-38

Lehrberger, J., "Automatic Translation and the Concept of Sublanguage", in Kittredge & Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, (de Gruyter, 1982), pp. 81-106

Liang, Y., "Vergleichende Darstellung von Fachtexten mit instruktiver Funktion im Deutschen und Chinesischen", in *Die Neueren Sprachen 87*, (1988, vol. 1/2), pp. 91-111

Lundquist, L., "Some Considerations on the Relations between Text Linguistics and the Study of Texts for Specific Purposes", in H. Schröder (ed.), *Subject-oriented Texts*, (Berlin: de Gruyter, 1991), pp.231-243.

Lyons, J., *Semantics*, (Cambridge & New York: Cambridge University Press, 1977)

Lyons, J., *Introduction to Theoretical Linguistics*, (London: Cambridge University Press, 1969)

Meijs, W. (ed.), "Corpus Linguistics and Beyond", *Proceedings of the 7th International Conference on English Language Research on Computerized Corpora*, (Amsterdam: Rodopi, 1987)

Milne, R., "Resolving Lexical Ambiguity in a Deterministic Parser", in *Computational Linguistics*, (Vol.12, No.1 1986). pp. 1-12

Morris, J. et al., "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", in *Computational Linguistics*, (Vol.17, No.1., 1991), pp. 21-48

Moskovich, W., "What is a Sublanguage: The Notion of Sublanguage in Modern Soviet Linguistics", in Kittredege & Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, (Berlin & New York: de Gruyter: 1982), pp. 191-205

Nelson, W., H. Hucera, *Frequency Analysis of English Usage: Lexicon and Grammar.*, (Boston: Miffin Co., 1982)

Nirenberg, S. (ed.), *Machine Translation-Theoretical and Methodological Issues,* (London: Cambridge University Press, 1987)

Nirenberg, S., *Machine Translation: A knowledge-based approach*, (San Mateo, California: Morgan Kaufmann, 1992)

Nordmann, M., "Rhythm and Balance in LSP Texts", in *Special Language - Fachsprache*, (Heft 1-2, 1989), pp. 24-36

O'Brien, S. & E. Quinlan, *Corpus Creation for Sublanguage Analysis*, (Eurotra Interim Report, May 1992)

O'Brien, S. & E. Quinlan, "Sublanguage: Characteristics and Selection Guidelines for MT", in *Proceedings of the 5th Annual Conference on AI and Cognitive Science*, (Limerick, Ireland, August, 1992), (forthcoming)

Oostdijk, N., *Corpus Linguistics and the Automatic Analysis of English*, (Amsterdam: Rodopi, 1987)

Opitz, K., "LSP versus Common Language: The Muddle of Definiens and Definiendum", in J. Hoedt et al (eds.), *Proceedings of the 3rd European Symposium on LSP: "Pragmatics and LSP"*, (Copenhagen, August, 1982), (Copenhagen: LSP Centre, UNESCO ALSED LSP Network and Newsletter, Copenhagen School of Economics, 1982), pp. 185-198

Pei, M.A., *Glossary of Linguistic Terminology*, (New York, Columbia University, 1966)

Philips, M., "Text Structure and Text Typology", in *Aspects of Text Structure*, (North Holland Linguistic Series, 1985), pp. 197-217

Quinlan, E. et al., *Sublanguage and Text Typology*, (Eurotra Final Report, August, 1990)

Quinlan, E. & S. O'Brien, *Sublanguage and Text Types*, (Eurotra Interim Report, November 1991)

Quinlan, E. & S.O'Brien, *Sublanguage Corpus Analysis: A Comparison of Two Domains*, (Eurotra Interim Report, August 1992)

Raskin, V., "Ontology, Sublanguage and Semantic Networks in NLP", in M Gloumbie (ed.), *Advances in AI*, (Springer Verlag, 1990), pp. 114-128

Renouf, Antoinette, "Corpus Development", in J. Sinclair (ed.), *Looking Up: An account of the COBUILD Project in Lexical Computing* ed. by J. Sinclair, (London: Collins ELT, 1987), pp. 1-40

Sager, J.C., "Machine Translation and a Typology of Texts", in Laurén & Nordmann (eds.): *Special Language: From Humans Thinking to Thinking Machines* (Papers presented at the 6th European Symposium on LSP, University of Vaasa, Aug. 3rd-7th, 1987), (Clevedon & Philadelphia: Multilingual Matters, 1989), pp. 397-410

Sager, J. et al., *English Special Languages*, (Wiesbaden: Brandstetterverlag, 1980)

Sager, N., *Natural Language Information Processing: A computer grammar of English and its applications*, (London & Amsterdam: Addison-Wesley, 1981)

Sager, N., "Syntactic Formatting of Science Information", in Kittredge & Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, (Berlin & New York: de Gruyter, 1982), pp. 9-26

Sager, N. et al., *Medical Language Processing: Computer Management of Narrative Data*, (USA & Canada: Addison-Wesley, 1987)

Salager-Meyer, F. et al, "Communicative Function and Grammatical Variations in Medical English Scholarly Papers: A Genre Analysis Study", in C. Laurén & M. Nordmann (eds.), *Special Language: From Humans Thinking to Thinking Machines* (Papers presented at the 6th European Symposium on LSP, University of Vaasa, Aug. 3rd-7th, 1987), (Clevedon &Philadelphia: Multilingual Matters, 1989), pp.151-160

Salton, G., *Automatic Text Processing*, (USA: Addison-Wesley, 1989)

Schröder, H., "Linguistic and Text-Theoretical Research on Languages for Special Purposes", in H. Schröder (ed.), *Subject-oriented Texts - Languages for Special Purposes and Text Theory*, (Berlin: de Gruyter, 1991), pp. 1-48.

Sebba, M., "The Adequacy of Corpora in MT", in *Applied Computer Translation*, (Vol.1, Issue 1, 1990), pp. 15-27

Slocum, J., *Machine Translation Systems*, (Cambridge, New York: Cambridge University Press, 1988)

Sowinski, B., *Textlinguistik*, (Stuttgart: Kohlkammer Verlag, 1983)

Spillner, B., "Textsorten im Sprachvergleich: Ansätze zu einer Konstrativen Textologie", in Kühlwein et al (eds.), *Kontrastive Linguistik und Übersetzungswissenschaft* (Akten des Internationalen Kolloquiums Trier/Saarbrücken, 25-29/9, 1978), (Munich: Fink, 1981), pp.239-249

Stolze, R., *Grundlagen der Textübersetzung*, (Heidelberg: Julius Groos Verlag, 1982)

Sumita, E. et al., "Experiments and Prospects of Example-Based Machine Translation", in *Proceedings of the 29th annual meeting for Computational Linguistics*, (1991), pp.185-192

Thomson, A.J., A.V. Martinet, *A Practical English Grammar*, (Oxford: Oxford University Press, 1986)

Trimble, L., *English for Science and Technology: A Discourse Approach*, (Cambridge: Cambridge University Press, 1985)

Ulijn, G., "Universals and Variants in Scientific and Technical English, French, German and Dutch", in J. Hoedt et al (eds.), *Proceedings of the 3rd European Symposium on LSP: "Pragmatics and LSP"*, (Copenhagen, August, 1981), pp.217-219

Van Dijk, T., *Grammars and Descriptions (Studies in Text Theory and Text Analysis)*, (Berlin & New York: de Gruyter, 1977)

Van Dijk, T. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*, (London & New York: Longman, 1977)

Van Dijk, T. (ed.), *Handbook of Discourse Analysis (Vol. 1)*, (London: Academic Press, 1985)

Van Dijk, T., "The Future of the Field: Discourse Analysis in the 1990s", in *Text* (vol. 10, 1990), pp. 133-156

Van der Eijk, P., "Linguistic analysis in the MiMo Translation System", in *Eurotra Working Papers Series,* Katkolieke Universiteit, Leuven, (1989)

Van den Eynde, K., "The Pronominal Approach in NLP - A pronominal feature analysis of co-ordination in French", in *Computers and Translation 3,* (Kluwer, 1988), pp.177-213

Van Eynde, F. et al, *Linguistic Issues in Machine Translation,* (London, New York: Printer Publishers, 1993)

Viehweger, D., "Methodologische Problem der Textlinguistik", in *Zeitschrift für Germanistik,* (Vol. 1, 1980), pp.6-20

Weber, H., "Language for Specific Purposes, Text Typology, and Text Analysis: aspects of a pragmatic-functional approach", in J. Hoedt et al. (eds.), *Proceedings of the 3rd European Symposium on LSP, "Pragmatics and LSP"*, (Copenhagen, August, 1982), pp. 219-234

Weber, H., *Converging Approaches in Machine Translation, Domain Knowledge and Discourse Knowledge*, Linguistic Agency, University of Duisberg, Series B, Paper no. 164.

Wirth, J. (ed.), "Beyond the Sentence: discourse and sentential form", in *Discourse, pragmatics and Linguistic Form*, (Karoma Publishers, 1985), pp. 1-19