SUBLANGUAGES IN MACHINE TRANSLATION

Heinz-Dirk Luckhardt
Fachrichtung 5.5 Informationswissenschaft
Universität des Saarlandes
D-6600 Saarbrücken, Federal Republic of Germany

ABSTRACT

There have been various attempts at using the sublanguage notion for disambiguation and the selection of target language equivalents in machine translation. In this paper a theoretical concept and its implementation in a real MT application are presented. Above this, means of linguistic engineering like weighting mechanisms are proposed.

INTRODUCTION

It has been proposed by a number of authors (cf. Kittredge 1987, Kittredge/Lehrberger 1982, Luckhardt 1984) to use the sublanguage notion for solving some of the notorious problems in machine translation (MT) such as disambiguation and selection of target language equivalents.

In the following, I shall give a rough summary of what sublanguages can contribute to the solution of concrete MT problems.

A SUBLANGUAGE CONCEPT FOR USE IN MT SYSTEMS

To my knowledge, it was Z. Harris who introduced the term 'sublanguage' (cf. Harris 1968, 152) for a portion of natural language differing from other portions of the same language syntactically and/or lexically. Definitions are given by Hirschman/Sager (1982), Quinlan (1989) and Lehrberger (1982).

In order to be able to use such characterizations in MT, they have to be formalized in a way adequate to the MT system in question. Such formalizable properties were combined in the definition of Luckhardt (1984) of what sublanguage can mean for MT:

Text type represents the syntactic-syntagmatic level of a sublanguage for which only a rather weak differentiation can be proposed (e.g. running text, word list, nominal structures etc.).

Subject field represents the lexical level of a sublanguage, i.e. for every sublanguage a subject field is determined as being characteristic, so that the MT system may choose on the basis of the sublanguage of a text those translation equivalents from the lexicon which carry the same subject field code as the translated text.

The lack of a commonly accepted subject field classification for MT is a serious problem. Such a classification is tentatively proposed in Luckhardt/Zimmermann 1991.

Text function represents the lexical-pragmatic level. The function of a text (or its target group) may determine the choice of TL equivalents and of syntactic structure or style.

The inhouse usage criterion covers a number of aspects determined by special requests of the MT user or the firm ordering the translation. This is first of all a question of inhouse terminology.

SUBLANGUAGES FOR MT: MAINTENANCE REQUIREMENTS

A typical maintenance requirement card of the Bundessprachenamt (Federal Translations Agency) among others contains the following parts:

1. designation of equipment
   text type 'nominal structure'
   text function 'title'
   e.g.: 'Portable gasoline driven pump'

2. tools, parts, materials
   text type 'word list'
   text function 'accessories'; e.g.:
   - key set, head screw, L-type hex
   - wrench, adjustable, open end 6"
   - solvent, type II
   - screwdriver, flat tip, medium duty
   - rags, wiping

3. procedure
text type 'instructions'
(imperative style)
text function 'maintenance
instructions', e.g.:

'Accomplish annually or when directed as a result of operational test. Clean and inspect fuel filter and float valve;
- remove pump housing covers, if applicable
- observe no smoking regulation
- remove choke knob and fuel connection
- remove float chamber and gasket
- clean all parts in solvent, allow to air dry
- inspect filter for clogging, tears, and deterioration'

(cf. Wilms 1983)

The example indicates how nicely the different sublanguages of this type of document can be differentiated, and it ought to be possible in all MT systems to capture these differences, especially the typical 'imperative style' of the text type 'instructions'. In order to achieve this it must be possible to weight rules or resulting structures like in the SUSY system (cf. Thiel 1987). This is important, because there is no absolute certainty that all predicate structures appear as imperatives in English or as infinitives in German.

THE USE OF SUBLANGUAGES IN THE STS PROJECT AND SYSTEM

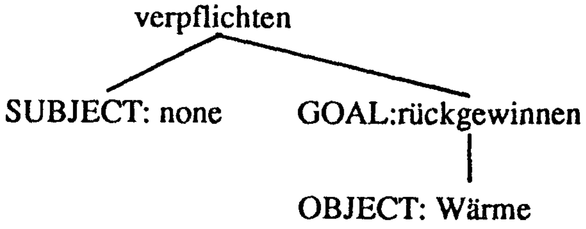
Since 1985 the SUSY system has been used as the core MT system within the computer-aided Saarbrücken Translation System (STS), i.e. in human-aided MT and in machine-aided human translation. Titles of scientific papers from German databases were machine-translated and postedited by humans, abstracts were translated by translators (in all around 5 million words), with the MT system automatically supplying the correct terminology (from a terminology pool of more than 350.000 German-English entries). In the following a specific aspect of sublanguage-dependent disambiguation is described.

SEMANTICS OF PREPOSITIONS IN TITLES

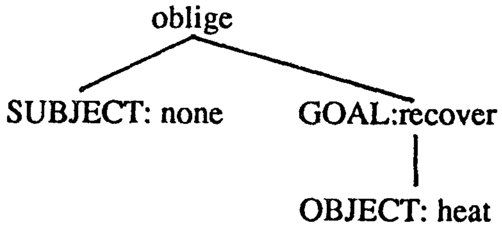
Highly ambiguous prepositions like 'zu', 'über' etc. can be safely disambiguated on the basis of word order:

'Zur Optimierung von Waldschadenserhebungen' => 'The optimization of wood damage surveys'
'Zur Rückgewinnung von Wärme verpflichtet' => 'Obliged to recover heat'
'Technologien zur Verminderung von Abfällen' => 'Technologies for the reduction of waste'
'Über Arbeit und Umwelt' => 'Labour and environment'

A 'zu'-phrase at the beginning of a title (the top node of the nominal structure) always denotes a TOPIC (1st example), otherwise (3rd example) a purpose. 'Über' at the beginning also denotes a TOPIC. These rules only apply, if the PP is not embedded in a predicate structure like in the 2nd example, where it fills the zu-valency of 'verpflichtet'. So, if the parser produces a structure like the following:



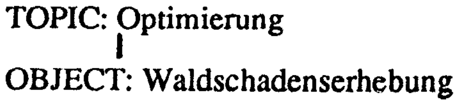
there only has to be lexical transfer =>



to present a structure to generation that carries enough information to produce the English translation given above ('Obliged to recover heat').

Similarly, examples 1. and 3. can be represented by the parser in a way which allows the generation of the correct target language equivalent, e.g.:

'Zur Optimierung von Waldschadenserhebungen'



transfer =>

TOPIC: optimization
|
OBJECT: wood damage survey

generation =>

'The optimization of wood damage surveys'

The surface realization of the semantic roles TOPIC and OBJECT is a task for generation, i.e. transfer can be completely relieved of rules treating such semantic roles (cf. Luckhardt 1987).

CONCLUSION

Sublanguage is a notion MT developers ought to turn their attention to

- when their system has reached a stable and robust state offering the necessary tools and methods of language engineering like weighting mechanisms
- when their system is about to be applied to large volumes of text with distinct sublanguage characteristics
- if a terminological data base system has been established which makes it possible to cover the lexical and inhouse usage levels of sublanguages and which can be accessed by the MT system
- if the necessary machine-readable terminology is at hand.

A sublanguage is not as easy to implement as it may appear from a first glance at texts of a specific corpus, however distinct that type of text may look. Very often the apparently formalizable criteria turn out to be useless for MT, although any human reader could easily formulate them. The METEO ideal of a sublanguage surely cannot be reproduced easily.

REFERENCES

Harris, Z. (1968). *Mathematical Structures of Language*. Wiley-Interscience

Hirschman, L.; N. Sager (1982). *Automatic information formatting of a medical sublanguage*. In: Kittredge/Lehrberger (eds., 1982)

Keil, G.C. (1982). *System Conception and Design. A Report on Software Development within the project SUSY-BSA*. Saarbrücken: Universität des Saarlandes: Projekt SUSY-BSA

Kittredge, R. (1987). *The Significance of Sublanguage for Automatic Translation*. In: S. Nirenburg (ed.). *Machine Translation. Theoretical and Methodological Issues* Cambridge University Press

Kittredge, R.; J. Lehrberger (ed., 1982). *Sublanguage. Studies of Language in Restricted Semantic Domain*. Berlin / New York

Lehrberger, J. (1982). *Automatic Translation and the Concept of Sublanguage*. In: Kittredge/Lehrberger (eds., 1982)

Luckhardt, H.-D. (1984). *Erste Überlegungen zur Verwendung des Sublanguage-Konzepts in SUSY*. In: Multilingua 3-3/1984

- (1987). *Der Transfer in der maschinellen Sprachübersetzung*. Tübingen: Niemeyer

- (1989a). *Terminologieerfassung und -nutzung im computergestützten Saarbrücker Translationssystem STS*. In: H.H. Zimmermann; H.-D. Luckhardt (eds., 1989). *Der computergestützte Saarbrücker Translationsservice STS*. Veröffentlichungen der FR 5.5 Informationswissenschaft. Saarbrücken

Luckhardt, H.-D.; H.H. Zimmermann (1991). *Computer-Aided and Machine Translation. Practical Applications and Applied Research*. Saarbrücken: AQ-Verlag

Quinlan, E. (1989). *Sublanguage and the relevance of sublanguage to MT*. Unpublished paper. EUROTRA-IRELAND. Dublin

Thiel, M. (1987). *Weighted Parsing*. In: L. Bolc (ed.). *Natural Language Parsing Systems*. Berlin: Springer

Wilms, F.-J. (1983). *SUSY-BSA: Abschlußdokumentation*. Teil I. Saarbrücken: Universität des Saarlandes: Projekt SUSY-BSA