DESIGN OF A MACHINE TRANSLATION SYSTEM FOR A SUBLANGUAGE

Beat Buchmann, Susan Warwick, Patrick Shann
Dalle Molle Institute for Semantic and Cognitive Studies
University of Geneva
Switzerland

ABSTRACT

This paper describes the design of a prototype machine translation system for a sublanguage of job advertisements. The design is based on the hypothesis that specialized linguistic subsystems may require special computational treatment and that therefore a relatively shallow analysis of the text may be sufficient for automatic translation of the sublanguage. This hypothesis and the desire to minimize computation in the transfer phase has led to the adoption of a flat tree representation of the linguistic data.

1. INTRODUCTION

The most promising results in computational linguistics and specifically in Machine Translation (MT) have been obtained where applications were limited to languages for special purposes and to restricted text types (Kittredge, Lehrberger, 1982). In light of these prospects, the prototype MT system described below[1] should be seen as an experiment in the computational treatment of a particular sublanguage. The project is meant to serve both as a didactic tool and as a vehicle for research in MT. The development of a large-scale operational system is not envisaged at present. The following research objectives have been defined for this project:
- to establish linguistic specifications of the sublanguage as a basis for automatic processing;
- to develop translation algorithms tailored to a computational treatment of the sublanguage.

The emphasis of the research lies in defining the depth of linguistic analysis necessary to adequately treat the complexity of the text type with a view to acceptable machine translation. It is the conjecture of our research group that, within the particular sublanguage defined by our corpus, acceptable translation does not necessarily depend on standard linguistic structural analysis but can be obtained with a relatively shallow analysis. Thus, as a working hypothesis, the principle of 'flat trees' has been adopted for the representation of the linguistic data. Flat trees, as opposed to deep trees, only partially reflect the dependency struc-

---

[1] Project sponsored by the Swiss government.

ture obtained by a traditional IC-analysis. The adoption of flat trees goes hand in hand with the further hypothesis that the sublanguage can be translated mechanically with only minimal semantic analysis similarly to the TAUM-METEO system (Chevalier, et al., 1978).

2. THE SUBLANGUAGE

The corpus is taken from a weekly publication by the Swiss government announcing federal job openings. The wordload of this publication amounts to ca. 10,000 words per week; however, many of the advertisements are carried for several weeks. All job adds are published in the three national languages: German, French and Italian, with German usually serving as the source language (SL), French and Italian as the target language (TL). The study is hence based on a collection of texts already translated by human translators. The ads are grouped according to profession, e.g. academic, technical, administrative, etc. At present, the corpus is limited to the domain of administrative positions, an example of which is given in figure 1.

Verwaltungsbeamtin
Fonctionnaire d'administration
Funzionaria amministrativa

Führen des Sekretariates eines Sektionschefs. Ausfertigen von Korrespondenzen und Berichten nach Diktat und Vorlage in deutscher, französischer und englischer Sprache. Abgeschlossene kaufmännische Lehre oder Handelsschulbildung. Berufserfahrung erwünscht. Sprachen: Deutsch, Französisch, Englisch in Wort und Schrift. Italienisch und/oder Spanisch erwünscht.

Diriger le secrétariat d'un chef de section. Dactylographier de la correspondance allemande, française et anglaise et des rapports sous dictée ou d'après manuscrits. Certificat d'employée de commerce ou diplôme d'une école de commerce. Expérience professionnelle désirée. Langues: le français, l'allemand et l'anglais parlés et écrits. Connaissances de l'italien ou de l'espagnol, voire des deux souhaitées.

Dirigere il segretariato di un capo sezione. Stesura di corrispondenza e rapporti secondo dettato o manoscritto. Tirocinio commerciale o formazione commerciale. Pratica pluriennale. Lingue: tedesco, francese, inglese (orale e scritto). Buone nozioni dell'italiano e/o dello spagnolo auspicate.

Figure 1. Advertisement for an administrative position ("Die Stelle", 1981).

The corpus exhibits many of the textual fea-
tures generally used to characterize a sublanguage,
i.e. (i) limited subject matter, (ii) lexical and
syntactic restrictions, and (iii) high frequency
of certain constructions. As can be seen from the
example, the style of the sublanguage is distin-
guished by complex nominal dependencies with va-
rious levels of coordination. In addition, most
sentences are incomplete in that they consist of a
series of nominal phrases and do not contain a main
verb; no relative phrases nor dependent clauses
occur. The importance of nominal constituents is
reflected in the statistics of the German texts:
over 55% of the words in the corpus are nouns,
11% adjectives, 11% prepositions, 17% conjunctions;
verbs only make up 1% of the corpus. A comparison
with the statistics of the French and Italian
translations reveal approximately the same distri-
bution except for infinitival verbs. The higher
frequency of verbs in French and Italian is due to
a preference for infinitival phrases in place of
deverbal nominal constructions. Apart from this
difference, the major textual characteristics
carry over from source to target sublanguage there-
by facilitating mechanical translation.

3. BRIEF DESCRIPTION OF THE SYSTEM

Modern transfer-based MT systems are based on
the following design principles : (i) modularity,
e.g. separation of linguistic data and algorithms,
(ii) multilinguality i.e. independent analysis,
transfer, and generation phases, (iii) formalized
specification of the linguistic model (Hutchins,
1982). Although only a prototype, the system was
designed in accordance with these considerations.

As to modularity, the software used is a gene-
ral purpose rule-based transducer especially deve-
loped for MT (Shann, Cochard, 1984). This software
tool not only allows for the separation of data
and algorithms but also provides great flexibility
in the organization of grammars and subgrammars,
and in the control of the computational processes
applied to them.

As a multilingual system it is not directly
oriented towards any specific language pair; the
same German analysis module serves as input for
the German-French as well as the German-Italian
transfer module. Separate French and Italian gene-
ration modules use only language specific knowledge
to produce the final translation. However, the Ger-
man analysis is indirectly influenced by target
language considerations: the interface structure
between analysis and transfer was defined to take
advantage of the similarities between the three
languages and to accommodate the differences.

4. LINGUISTIC APPROACH: MINIMAL BUT SUFFICIENT
    DEPTH

With the sublanguage investigated displaying
restricted syntactic structures within a limited
semantic domain, a grammar specifically tailored to
these job advertisements can be defined. Moreover,

the linear series of nominal phrases as well as
the almost one-to-one lexical equivalences found
in the SL and TL texts suggest that a shallow ana-
lysis without a semantic component is sufficient
for adequate translation. The flat tree represen-
tation resulting from such a minimal depth approach
does not make any claim to linguistic generaliza-
bility for purposes other than the translation of
this particular sublanguage.

4.1 Computational considerations

In a transfer-based MT system, actual trans-
lation takes place in transfer and can be descri-
bed as the computational manipulation of tree
structures. In the absence of any formal theory of
translation for MT, and given the relatively well-
developed analysis techniques currently available,
a major concern in MT research is to minimize the
computation necessary in the transfer phase. A
flat tree representation provides one way of sim-
plifying the structures to be processed; an inter-
face representation defined to accommodate both
SL and TL structures in the same manner, thus
avoiding tree structure manipulation, is yet ano-
ther means. The representation of the linguistic
data in this system is a direct result of these
two considerations.

4.2 Flat trees

The fact that the linearity of the surface
structure constituents carries over from SL to the
TLs justifies the adoption of a minimal depth ana-
lysis. The analysis is restricted to the identifi-
cation of the phrasal constituents and their inter-
nal structure; dependencies holding between consti-
tuents are only partially computed. Thus, the
interface structure resulting from analysis and
serving as input to transfer does not reflect a
linguistically correct dependency structure.
Instead, the IS respects the linear surface order
of the constituents (with the exception of predi-
cate groups, see below) in a flat tree represen-
tation.

In a flat tree, the major phrasal consti-
tuents, in particular the prepositional phrases,
are not attached at the node from which they de-
pend linguistically but at specified nodes higher
up in the tree. Schematically, the differences
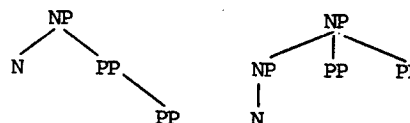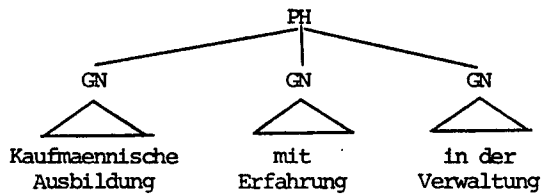can be illustrated as follows:

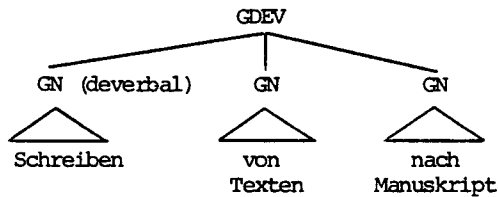Fig. 2. Standard IC-tree vs. Flat tree

The flat tree representation applies to all three
major phrasal constituents defined for this cor-
pus: (i) nominal phrases proper, (ii) deverbal

335

nominal phrases, and (iii) verbal phrases. Samples taken from the corpus are given below to illustrate each of the three constituent structures.
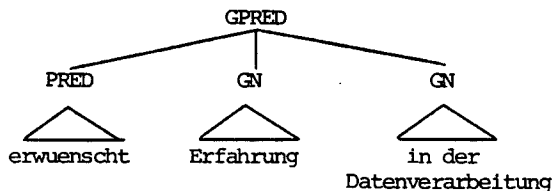
(i) Nominal phrases proper have a standard noun phrase as their head, possibly followed by a linear sequence of prepositional phrases. (GN stands for both standard NPs and PPs.)

```
                    PH
        _____/____|_____
       GN          GN          GN
      /__\        /__\        /__\
  Kaufmaennische   mit        in der
   Ausbildung    Erfahrung   Verwaltung
```

(ii) Deverbal nominal phrases have a deverbal noun as their head, followed by a linear sequence of GNs.

```
                  GDEV
        _____/____|_____
   GN (deverbal)   GN          GN
      /__\        /__\        /__\
   Schreiben      von         nach
                 Texten     Manuskript
```

(iii) Verbal phrases have a predicate as their head, followed by a linear sequence of GNs. (PRED encompasses predicative participles, predicative adjectives, and infinitival predicates; the few finite verbs in the corpus (0.4%) are not treated.)

```
                  GPRED
        _____/____|_____
     PRED          GN          GN
      /__\        /__\        /__\
  erwuenscht    Erfahrung    in der
                           Datenverarbeitung
```

("Erfahrung in der Datenverarbeitung erwuenscht")

4.3 Normalized tree structures

In order to further minimize manipulation of structure in transfer, the interface representation is also normalized for two important categories in the sublanguage, namely deverbal nominal phrases (GDEV) and noun and prepositional phrases (GN). The structures are defined such that they remain valid for both the source and target language.

4.3.1 Deverbal nominal phrases

A marked stylistic difference between the SL and the TLs occurring with high frequency in the corpus is the translation of a German deverbal noun into an infinitive in French and Italian. With the deverbal noun in German usually serving as the head of a complex nominal structure with several complements, the translation of the noun into an infini-.

tive in the target language changes the type of complement structure accordingly. The complete linearization of the deverbal complements provides a format for accommodating the target language infinitival construction aimed at in translation. Structural transfer is thus reduced to renaming the nodes; the normalized tree structure remains the same, as can be seen in the SL and TL representations shown below.
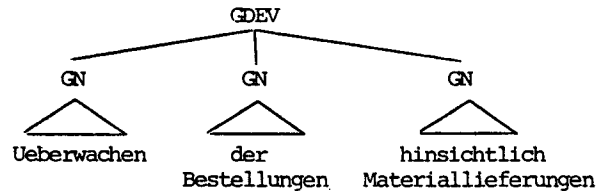
```
                  GDEV
        _____/____|_____
     GN           GN          GN
      /__\        /__\        /__\
  Ueberwachen     der        hinsichtlich
                Bestellungen Materiallieferungen
```

Fig. 3. SL (German) deverbal nominal phrase analysis.

```
                  GPRED
        _____/____|_____
    PRED           GN          GN
      /__\        /__\        /__\
  Surveiller      les        quant à la
                commandes   livraison du materiel
```
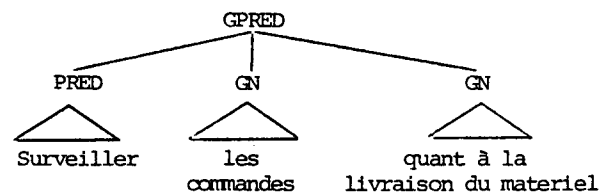
Fig. 4. Equivalent TL (French) verbal phrase analysis.

4.3.2 Noun phrases and prepositional phrases

Certain noun phrases in German (e.g. genetive attributes) are translated into prepositional phrases in French and Italian. In order to avoid structural transfer of noun phrases into prepositional phrases and vice-versa, a normalized form for noun phrases has been defined which reserves a position in the tree for prepositions. For standard noun phrases a special value (NIL) has been defined to fill the empty preposition slot. Therefore, in the transfer phase, a translation from a noun phrase to a prepositional phrase or vice-versa is merely a change in the value of the prepositional slot without any change in the tree structure.
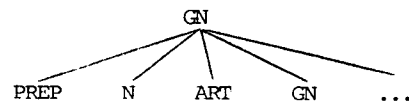
```
                  GN
        ____/___/__|_____
     PREP    N    ART    GN    ...
```

Fig. 5. Example of the normalized form for NPs and PPs.

4.4 CONSIDERATIONS FOR TRANSLATION

The goal of the system, and perhaps of MT in general, has to be to carry over the information content from SL to TL, to produce output acceptable

in terms of TL conventions, and to respect the
style of the text type. It seems that treating a
well-defined sublanguage enhances the possibili-
ties for an MT system to answer these requirements.
In fact, the sublanguage itself suggests possible
strategies for dealing with some of the classical
translation problems in MT such as (1) lexical
ambiguity, (2) translation of prepositions, and
(3) treatment of coordination.

4.4.1 Lexical problems

Two well-known lexical problems in computatio-
nal linguistics are homograph resolution and poly-
semy disambiguation. Given the small number of
possible syntactic structures in the sublanguage,
the few homographs found in the corpus do not pre-
sent any problems for analysis. In turn, the limi-
ted semantic domain of the sublanguage completely
eliminates multiple word senses so that the trans-
fer of lexical meanings is basically a one-to-one
mapping. Therefore, with the nouns serving as the
major carriers of the textual meaning, lexical
transfer ensures that the information content of
the text is carried over.

4.4.2 Translation of prepositions

The fact that the types of nouns occurring in
the sublanguage are restricted and repetitive and
that the possible prepositions commanded by any
given noun is small in number (max. 3 in the cor-
pus) allows the adoption of a limited noun-focused
approach for the translation of prepositions. In
such an approach, it is the particular noun or
noun class rather than general semantic features
that determine the translation of prepositions.
At present, the information relevant to correct
translation of prepositions is attached to indi-
vidual noun entries in the transfer dictionary;
semantic noun subclassification similar to other
sublanguage research (Sager, 1982) is being
investigated.

4.4.3 Coordination

With SL and TLs exhibiting parallel surface
syntactic structure, and with inherent ambiguities
of scope therefore carrying over, analysis of co-
ordination remains shallow. Conjunctions and in-
trasentential punctuation are defined functionally
as coordinators to yield, in keeping with the flat
tree representation, a structure such as the one
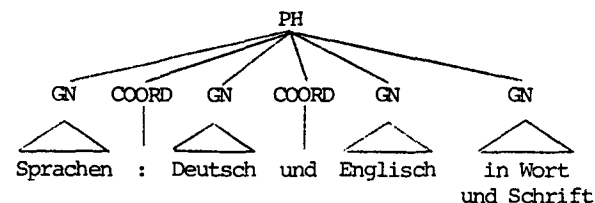shown below.

Fig. 6. Coordinated structure at sentence level.

5. CONCLUSION

The evidence available to-date seems to show
that, for the particular sublanguage dealt with,
correct translation is feasible under the hypo-
theses described in this paper. The non-generali-
zability of such an approach is quite evident;
however, the fact that such a 'minimal depth' ap-
proach seems to work for this particular sublan-
guage gives substance to the impression that spe-
cialized linguistic subsystems differ quite
sharply, both in complexity and linguistic fea-
tures, from the standard language and may there-
fore require special computational treatment.

REFERENCES

Chevalier et al. TAUM-METEO, Description du sys-
tème. Université de Montréal, 1978.

Eidgenössisches Personalamt (ed.). Die Stelle.
Stellenzeiger des Bundes. No. 21, 1981.

Grischman, R., Hirschman, L. and Friedman, C.
"Natural Language Interfaces Using Limited
Semantic Information." Proc. 9th International
Conference on Computational Linguistics, 1982.

Hutchins, W.J. "The Evolution of Machine Transla-
tion Systems." In: Lawson, V. (ed.), Practical
Experience of Machine Translation, Amsterdam,
N.Y., Oxford, 1982.

Kittredge, R., Lehrberger, J. (eds.). Sublangua-
ges, Studies of Language in Restricted Do-
mains, Berlin, N.Y., 1982.

Sager, N. "Syntactic Formatting of Science Infor-
mation." In: Kittredge, Lehrburger, 1982.

Shann, P., Cochard, J.L. "GTT : A General Trans-
ducer for Teaching Computational Linguistics."
COLING Communication, 1984.