Implicitness as a Guiding Principle
In Machine Translation

Klaus Schubert
RISO/Research, Postbus 8348, NV 3503 RH Utrecht, The Netherlands
schubert@r11.riso

Abstract

Bilingual extensibility requires an MT system to have a language-independent pivot. It is argued that an ideal, purely semantic pivot is impossible. A translation method is described in which semantic elements are kept implicit in syntax, while the semantic units and distinctions are explicit in the words of a full-fledged language used as pivot.

1. Multilingual extensibility

There is an external factor with very substantial consequences for the internal design of machine translation systems: extensibility. When a machine translation system has to allow for adding arbitrary source and target languages without each time adapting the already existing parts of the system, the need arises for a carefully defined interface structure to which modules for additional languages may be linked. The design that best meets these requirements is the pivot or interlingual approach, since in such a system there is only a single interface which gives access to all the languages already included in the system.

In models of this type the only link between a source and a target language is the intermediate representation. It has a double function:

1. The intermediate representation should render the full content of the text being translated, with all its details and nuances.
2. The intermediate representation should contain the results of the grammatical analysis carried out on the source text, where these characteristics are translation-relevant.

It is desirable that the intermediate representation express both the content and the grammatical characteristics of the text unambiguously, and since it is the interface to arbitrary languages, it should express them in a language-independent way.

2. Language-independent semantics?

To render both the content and the functional features of a text in such a way that they can be processed, the intermediate representation provides a formalism for this purpose. Spelling out means making explicit. My main concern here is investigating to what extent the required explicitness can be achieved in a language-independent representation. Are there language-independently valid categories and values for the characteristics of words and word groups needed in an intermediate representation? (When speaking of grammatical analysis, I take grammar to denote the study of the entire internal system of language, so that both syntax and semantics on all levels between morpheme and text are subfields of grammar. Pragmatics, by contrast, describes the influence of extralinguistic factors on language and is not part of grammar; cf. Schubert 1987b: 14f.)

The form of the linguistic sign is language-specific, whereas its content is normally thought to be language-independent. The content side of the linguistic sign is therefore often assumed to be a good target for translation grammar. In other words, the transfer step from a syntactic form in the source language to a corresponding form in the target language is performed on the basis of the common meaning the two forms are supposed to have.

As a consequence, an intermediate representation is usually devised as a structure in which this common meaning is made explicit. The intermediate representation is seen as a semantic equivalent of the source text. For obtaining such a structure, a syntactic analysis of the source text is by no means superfluous. An intermediate representation consists, like any system, of elements and their relations. In a semantic system elements and relations are semantic. But in order to detect the elements and their relations in a given text, a syntactic analysis is needed. ("Syntax-free semantic parsers" apply syntactic knowledge tacitly, and as a rule they work especially well for languages where the sequential order of "purely semantic" elements carries syntactic information.)

There are two major clusters of reasons why an ideal semantic intermediate representation of the language-independent kind sketched above is impossible, however desirable it may be in theory.

First of all, there are no language-independent semantic elements. Whatever symbols are chosen — words, morphemes, numbers, letter codes... they are always inherently language-bound. The elements of an artificial symbol system are either directly taken from an existing language, or have an explicit or implicit definition in a reference language. It is impossible to make a truly language-independent system of symbols, if it is to possess the full expressiveness of a human language (cf. Schubert 1986). Symbols cannot be given a meaning independently of a reference language, their meaning can only become autonomous by being used in a language community during a long period. This is why a planned language like Esperanto could not rank as a full-fledged human language from the very day the first textbook was published but had to develop slowly from an artificial, reference-language-dependent symbol system into an autonomous language by being used in a community (cf. Schubert forthc.). Perhaps this is an unusual argument in a computational context, where people are used to defining symbol systems which they call "languages". It should be borne in mind, however, that such defined symbol systems are subsets of an existing human language (or of several). Machine translation, by contrast, is concerned with translating texts between human languages, which from a semantic point of view — even if the language may be simplified or the text pre-edited — are inherently more complicated than artificial symbol systems.

Not only are defined semantic units in such systems reference-language-dependent, but the road to the basic semantic units needed is via semantic decomposition — with all its well-
-known problems. Scholars have for centuries been trying to find universally valid semantic atoms (or primitives), but none of the many systems suggested has met with acknowledge-

ment or proved applicable on any wider scale. Individual languages cut up and label reality in different ways; no under-
lying "smallest semantic units" have been found as yet and possibly they will never be found. In my opinion the conclu-
sion is that meaning is not portioned, so that no smallest
portions can be found.

Semantic atoms would be needed for totally spelling out the content of a text in a language-independent way, that is, in such a way that it would be suited for translation into any ar-
bitrary target language. In many machine translation systems, ambitions are not that high. Most often, intermediate representa-
tions use words or other language-bound symbols, decorated with semantic features which are held to be cross-
linguistically valid. Yet, what is true for semantic atoms ap-
plies to semantic features as well, albeit in a less obvious way: They contain portions of meaning which do not function in all languages in the same way. That semantic atoms and features are not as cross-linguistic as they seem to be, is also suggested by the experience that they are very hard to define and delimit in a way that fulfills exactly the required function, or denotes precisely the intended distinction for a large number of languages simultaneously. It is because of this that intermediate representations often have to be adapted, attuned or even redesigned when a new source or target language is added to the system. Such representations fail to provide for multilingual extensibility.

3. Case frames

The second cluster of reasons for the impossibility of an ideal, purely semantic, intermediate representation concerns seman-
tic relations. One of the best-known approaches to making semantic relations explicit is Fillmore's case grammar (1968). Deep cases are often believed to be cross-linguistically valid. Although there are many substantial difficulties in del-
imiting and labelling deep cases (cf. Fillmore 1987), many machine translation systems perform transfer with case frames. This works quite well to a certain degree, but slowly the insight is gaining ground that deep cases nevertheless are language-specific. If case frames really were an autonomous tertium comparationis, translating on the basis of case frames would mean just filling in target language forms in a language-independent case frame obtained from the source language analysis. But in reality case frame-based translation often entails a transfer from a source language-specific case frame to a target language one. Evidence for this need comes first from general linguistics (e.g. Pleines 1978: 572; Engel 1980: 11), but recently turns up in computational linguistics as well (Tsuji 1986: 656; cf. Schubert 1987a). This is in con-
ductor with Harold Somers' (1987: viii) observation about the popularity of case grammar, already declining in theoretical linguistics, but still in vogue in computational applications.

Returning to the argument about a purely semantic system, it can be concluded that neither the elements nor the relations, which together should constitute the theoretically desirable language-independent intermediate representation, actually ex-
ist. This insight, among others, is the origin of the idea of im-
pleness in machine translation.

4. Implicitness

Since there are no cross-linguistically valid semantic relations, and since case frames are therefore language-specific, the transfer step actually lacks a language-independent intermedi-
ate stage. This means that, where semantic relations are con-
cerned, there is no true pivot. There are only source structures and target structures with a transfer step somewhere between them. Given the notorious difficulties of defining deep cases, the question arises whether it is really necessary for machine translation to make semantic relations explicit. As they are language-specific anyway, it is much easier to perform transfer at another level, which is language-specific as well, but about which there is much more certainty: syntax. If transfer is carried out at the syntactic level, semantic deep cases can remain implicit.

Before describing this in somewhat more detail, a few words about the semantic elements. If there are no language-

independent semantic relations, looking for language-

independent semantic elements does not seem worthwhile ei-
ther. Yet, the above discussion of the function of an inter-
mediate representation entails another unexpected implication: Since an intermediate representation is the only link between source and target languages, it must be as expressive as any of them. If high-quality machine translation is the goal, this condition is inevitable, since the intermediate representation has to render and to convey the full and unsimplified content of the text, to make further translation possible. It must be feasible to translate into such an intermediate representation from all other languages. Interestingly enough, this translata-

tility criterion is the property by which human language is distinguished from artificial symbol systems by one of the classics of linguistics, Louis Hjelmslev (1963: 101). Accord-
ing to him, a human language (his term is daglagsprog) is a language into which all other communication systems (human languages and artificial symbol systems) can be translated. As a consequence of Hjelmslev's theory, an intermediate representation with the expressiveness indispensable for mul-
tilingual high-quality machine translation should indeed be it-
self a human language.

Now the elements and relations in the semantic system of the intermediate representation can be considered together. The discussion so far has yielded two results: There are no language-independent semantic elements and there are no cross-linguistically valid semantic relations. Moreover, the re-
quired expressiveness entails the consequence that the inter-
mediate representation should be a full-fledged language.

If the pivot of a machine translation system is a language (rather than an artificial symbol system), this removes the problems of spelling out semantic elements and relations. Se-
manics can then be kept implicit, that is, it can be expressed in the intermediate language by purely linguistics means, in the way illustrated below.

If the intermediate language is a full language, the syntactic side of the translation process comes down to performing two direct translations: first from a source language into the inter-
mediate language, and then from the intermediate into a target language. Moreover, if one opts for a human intermediate language, this brings about a substantial change in the design of a pivot-based multilingual machine translation system. Ar-
tificial intermediate representations are designed to achieve multilingual extensibility at the level of transfer. The condi-
tions that provide for extensibility are thus directly intertwined with the mechanisms that translate from one particular language into another. But when the intermediate representa-
tion is a language, multilingual extensibility shifts to another level: it is now catered for by the combination of language pair modules in which the intermediate language is always one of the two counterparts. This considerably facilitates the design, since multilingual extensibility with all its needs of cross-linguistically valid grammatical elements and relations no longer interferes with the translation steps proper. For this type of direct translation within a language pair, a translation method that performs the syntactic transfer on the basis of syntactic functions is both suitable and sufficient.

A possible implementation of this idea is found in the meta-
taxis translation method (Schubert 1987b: 222ff.). It works on the basis of language-specific syntactic functions and contras-
tive transformation rules that cater for the transfer step. Metat-
axis rules can be seen as contrastive lexical redundancy
rules over a bilingual dictionary. Technically speaking, they
are tree transduction rules which presuppose the dictionary to
consist of tree-structured entries. Metataxis is contrastive
dependency syntax for translation. Of course it is not the only
possible way of performing the syntactic part of a machine
translation procedure. A dependency-based approach, how-
ever, is especially well suited for a multilingual system, since
dependency syntax takes syntactic functions as its primary
units, using syntactic form as a secondary means. This is an
essential enhancement, since syntactic functions – i.e. depen-
dency relations such as subject, object etc. – are translation-
relevant, whereas syntactic form characteristics – such as a
word’s position vis-à-vis other words, its endings for case,
number, person, tense, mood, aspect etc. – are needed for
monolingual analysis and synthesis steps in an overall transla-
tion process, but are not themselves directly translation-
relevant).

As for the semantic side of the translation process, an in-
mediate representation represents its designers to make explicit all
the semantic distinctions needed for specific source and target
languages, which ultimately leads astray if multilingual exten-
sibility is aimed at. This is the danger of an “exploding” pivot.
If the pivot is a language, the degree of semantic detail it pro-
vides can be taken as a natural limitation to this explosive ten-
dency: An implementation is possible in which the entire se-
matic processing needed for a machine translation procedure
is carried out with linguistic means in the intermediate
language only. This means that whatever semantic elements or
relations are used, they are always expressed by means of
words and morphemes from the intermediate language. No se-
matic features, no selection rules and no meta-linguistic la-
beis or tags are used. This is in good agreement with the
metataxis approach to the syntactic side of the process: Meta-
taxis provides all syntactically possible translations of a source
sentence (clause, paragraph ...) and the semantic processing
performs a choice among these alternatives. (It normally needs
a substantial pragmatic augmentation with knowledge of the
world etc.; cf. Papegaaij/Schubert forthcoming chapter 3.5.). This
semantic process can be carried out entirely in the intermedia-
te language and is thus suitable for metataxis alternative
translations generated from whatever source language.

The second half of the translation, from the intermediate into
a target language, could in theory work in the same way, but
this would presuppose semantic processing in all the different
target languages. The requirement of extensibility is much
better met, if all the semantic processing for the second half
as well is carried out by means of the intermediate language.
This is indeed possible. The semantic-pragmatic processing in
the second half is – to put it in plain words – concerned with
fitting in the alternative translations offered in the bilingual
dictionary (intermediate language → target language) into the
context of the sentence and the entire text. What is needed for
assessing the probability of different contexts is information
about the typical contexts of the words in question: word ex-
pert knowledge. It is possible to describe the typical contexts
of target language words by means of words and phrases in the
intermediate language. Thus all semantic-pragmatic com-
parisons and probability computations are carried out ex-
clusively in the intermediate language, and as a consequence
only a single semantic system is needed for translating
between arbitrary languages: a system in the Intermediate
language. If this central system is built up within the limita-
tions of the intermediate language without reference to any
peculiarities of particular source and target languages, the
requirement of complete extensibility is fulfilled.

5. Conclusion
An intermediate language for high-quality machine translation
needs to be a full-fledged human language, due to the inherent
lack of expressiveness that is an inevitable characteristic of
artificial symbol systems. I argue that one can make a virtue
of this necessity: A human language as intermediate representa-
tion allows for rendering the full content of the text without
making semantic elements and relations more explicit than
what is expressed by appropriately interrelated words of the
intermediate language.

Of course the question arises whether, in that case, any arbi-
trary language would be suited for this function. It should be
pointed out, however, that the full range of trade-offs related
to the choice of an intermediate language cannot be dealt with
in this three-page contribution. My ideas about implicitness
are closely related to one of at least three fundamental criteria
for an intermediate language: expressiveness. The other two
are regularity and semantic autonomy. Only when all cri-
teria are considered together, can a choice be made.

References
Engel, Ulrich (1980): Fügungspotenz und Sprachvergleich. Vom
Nutzer eines semantisch orientierten Valenzbegriffs für die
konstruktive Linguistik. In: *Wirkendes Wort 30*, pp. 1-22
Fillmore, Charles J. (1968): The case for case. In: *Universals in
linguistic theory*. E. Bach / R. T. Harms (eds.). New York:
 Holt Rinehart & Winston, pp. 1-88
Fillmore, Charles J. (1987): A private history of the concept
"frame". In: *Concepts of case*. René Dirven / Günter Radden
(eds.). Tübingen: Narr, pp. 28-36
Hjelmslev, Louis (1963): Sproget. København: Berlingske forlag
[2nd ed.]
Papegaaij, B. C. / Klaus Schubert (forthc.): Text coherenee in trans-
lation. Dordrecht/Providenee: Foris
Pleines, Jochen (1986): Linguistic and extra-linguistic knowledge.
In: *Computers and Translation 1*, pp. 125-152
Schubert, Klaus (1987a): Wann bedeuten zwei Wörter dasselbe?
Über Tiefenklausus als Tertium comparationis. In: *Linguistik in
Deutschland*. Werner Abraham / Rivka Atthamir (eds.).
Tübingen: Niemeyer, pp. 109-117
Schubert, Klaus (1987b): Meataxis. Contrastive dependency syntax
for machine translation. Dordrecht / Providence: Foris
Schubert, Klaus (forthc.): Ausdruckskraft und Regelmäßigkei.
In: *Language Problems and Language Planning 12* [1998]
Somers, H. L. (1987): Valency and case in computational linguis-
tics. Edinburgh: Edinburgh University Press
Tsujii, Jun-ichi (1986): Future directions of machine translation.
In: *11th International Conference on Computational Linguis-
tics, Proceedings of Coling ’86*. Ronen: Institut für
angewandte Kommunikations- und Sprachforsehung, pp. 655-
668