Particle Homonymy and Machine Translation

Károly Fábricz

JATE University of Szeged,

Egyetem u. 2.

Hungary   H - 6722

Abstract

The purpose of this contribution is to formulate ways in which the homonymy of so-called 'Modal Particles' and their etymons can be handled. Our aim is to show that not only a strategy for this type of homonymy can be worked out, but also a formalization of information beyond propositional content can be introduced with a view to its MT application.

1. Introduction

During the almost 40 years of its existence machine translation has undergone a considerable refinement in the fields of both syntactic parsing and semantic representation. The development of MT can be seen as a tendency to incorporate more and more linguistic knowledge into the formalization of translational processes. Formalization has thus become a keyword for MT and has had several major implications. Firstly, it refers to the hypothesis that everything related to a given language is s t r u c t u r e d in one way or another. Secondly, formalization is an o b j e c t i v e m e a n s of testing the validity of the linguist's hypotheses about linguistic phenomena. Thirdly, it involves the linguist's h o p e that anything that has to do with language can in fact be formalized.

At present, there are several semantic theories which could be labelled "formal semantics". They are preoccupied with exploring the propositional content of different text-units and they do not deal with the phenomenon of "subjectivity". Subjectivity, or self-expression, as Lyons /1981, 240/ has pointed out, "cannot be reduced to the expression of propositional knowledge and beliefs". If we think of MT in its ideal form, i.e. not as an abstracting device, but as a system producing automatic translation, then the inadequacy of restriction to propositional content will be evident.

The present paper sets out to show that the expression of lexical subjectivity, conveyed by modal particles, should, and can, be accounted for in the process of MT.

2. Particle Homonymy

Let us consider the following pairs of sentences:

/1/ a. *There is <u>only</u> a little beer left.*
   b. *I was <u>only</u> too pleased to leave that place.*

/2/ a. *<u>Nur</u> ihn hatte man vergessen.*
   b. *Wozu habe ich <u>nur</u> gelebt?*

/3/ a. *Vous partez <u>déja</u>?*
   b. *Comment vous vous appelez <u>déja</u>?*

/4/ a. *Приходите к нам <u>и</u> завтра.*
   b. *Я <u>и</u> не знаю, что сказать.*

/5/ a. *Anna <u>is</u> eljön hozzánk?*
   b. *Hol <u>is</u> tartottunk?*

The words underlined in the b. example of each pair of sentences belong to a word-group now more or less uniformly referred to as 'Modal Particles' /cf. Arndt 1960/.

These words represent, in Arndt's term, a grammatical no-man's-land, although in the past ten years there has been a considerable interest towards modal particles.

Words like the English *only* or the German *nur* present two problems from the point of view of machine translation. On the one hand, they are ambiguous and their homonymy must be resolved. On the other hand, when such lexemes are used as modal particles, their "translation" causes serious problems since we can rarely translate the modal *only* into German as *nur*, or, say, into Hungarian as *csak*.

3. Resolution of Homonymy

As far as homonymy is concerned, clearly the task is to set up formal rules for the categorization of a given word as opposed to its alternative morphological and syntactic status.

The implication of the assignation of such homonymous lexemes to certain classes of words is by no means a matter of "simple" selection restriction at surface level. Each modal particle has preserved much of its etymon's syntactic and semantic properties. Given this, it follows that the ambiguity may be resolved by constructing small "subgrammars" for each of these particles, so as not only to set them apart from their homonyms, but also to take into consideration the whole communicative content of the sentence.

Thus, a subgrammar recognizing *only* — either as a logical operator, with its restrictive meaning, or as a modal particle, with its vague and, in a sense, antonymous meaning — would have to be capable of manipulating information from different levels. By comparing sentences /1a/ and /1b/ it could be concluded that, say, *only* is an operator when it precedes an NP /e.g. Det + Adj + N/ and is a particle when followed by *too*. But this assumption can readily be proved faulty by considering /6/:

/6/ *If only you had come, you could have saved me a lot of trouble.*

It is commonly held that, in order to parse sentences, one needs strategies for locating verbs and their complements, assigning words to various categories, depending on context /Lehrberger 1982, 102/. The recognition of particles can be done mainly by starting from semantic representations which should contain information concerning both the propositional content of sentences and their extrapropositional, or subjective modal content. Thus, assigning *only* to particles would imply an algorithm roughly defined as: "If the lexeme *only* is used with a word that has no restrictive component in its meaning, then it is a particle; otherwise it is an operator".

Parsing along these lines would mean a very complicated presentation of different parts of speech, including not only NPs, made up of adjectives, nouns, but also adverbs, pronouns and even phrases to account for *only* constructions like /6/. In addition, a very sophisticated and precise definition of the restriction/non-restriction opposition would have to be set up.

Obviously, the difficulty of assigning homonymous lexemes to modal particles, on the one hand, and to operators, intensifiers, adverbs,conjunctions, and the like, on the other, lies in the fact that the former bear a relationship to the overall meaning of the sentence, while the latter add their meaning to the global meaning only via some lower level of semantic structure.

From the above consideration it follows that it would be a fairly tedious and probably unreasonable task to attempt to resolve this kind of homonymy by the algorithmization of abstract sense-components.

Instead, it might be sufficient to construct a subgrammar to check *only* and other homonyms solely by reason of their being a particle. One way to make the information contained in the subgrammar available to the parser may be to indicate, in the dictionary entry of the homonym, all the cases in which the given word could possibly appear as a particle.

In English, or French, the resolution of ambiguity would mean setting up as few as 6-10 subgrammars, while in German, Russian

or Hungarian there are scores of homonymous particles and, consequently, subgrammars. In addition, the latter languages make quite frequent use of particle combinations which do not, as a rule, derive their meanings from a complement of the two /or more/ particles, but have some different meaning, cf.

/7/ *Csak nem fáztál meg?*

/8/ *Уж не простудился ли ты?*

Nevertheless, there seems to be no reason why these combinations could not be included in the subgrammar under one or the other dictionary entry.

4. Translation of Modal Particles

Whilst intensifiers, conjunctions, operators, pronouns, or adverbs have meanings which may be considered more or less "universal", the semantics of particles takes us into a field specific to a particular language. In other words, using *only* as an operator is "almost" identical to using, say, *nur*, or *seulment*, or *csak* etc. as an operator in German, French or Hungarian respectively. But when it comes to particles, we may experience difficulties in preserving the operator equivalent of *only* in the translation of sentences like /1b/ into any other language.

One possible solution, as with lots of different types of translation, would simply be to consider these words irrelevant from the point of view of propositional content matching. However, it would seem more plausible to try to find equivalents to these particles in the target language since, depending on the type of context to be translated, the expression of subjectivity may play a major role in producing the actual communicative message.

Functional equivalence is a notion frequently used in linguistic theory /Arnol'd 1976; Sanders 1980/, and it can be applied as a yardstick in particle matching /Fig. 1/. A study of modal particle translation is now being undertaken in Szeged University's Eng-

lish-Hungarian MT project and it is based on functional equivalence.

Those researchers who study MT in restricted semantic domains might overlook the problem of the subjectivity of the different texts. It should be noted, however, that "most of the unexpected structures one finds in a sublanguage text can be associated not so much with a shift in semantic domain as with a shift /usually quite temporary/ in the attitude which the text producer takes towards his domain of discourse" /Kittredge 1982, 135/. But even with academic papers it happens to be the case that during their translation one should be aware of the appearance of some subjective overtone lest some mistranslation should ensue. In this respect, consider the following two examples with *only* as a particle:

/9/ *Only too often have far-reaching conclusions been drawn from inadequate data collected from a limited number of languages.* /Ullmann:Semantic Universals, 1966, p. 218/

/10/ *Similarly, it is only natural that verbs for "snoring" should in many languages contain an /r/...* /Op. cit. p. 225/

The foregoing considerations lead us to the following sketchy representation of *only*:

```
ONLY - MP if - preceded by if + optative     A
             - followed by too + adverb/      B
                                   adject.
             - followed by ADJ                C
        - else  - LO                          D
     A translates as  a: Bárcsak + cond. if
                          Simple Sentence
                       b: Ha + ∅ + cond. if
                          Complex Sentence
     B translates as  is
     C translates as  csak
     D translates as  csak
```

Fig. 1. Subgrammar of *only* based on its Hungarian functional equivalents MP = Modal Part. LO = Log. operator