Requirements on Linguistic Knowledge Sources for Multilingual Generation

Johannes Matiasek and Harald Trost
Austrian Research Institute for Artificial Intelligence*
Schottengasse 3, A-1010 Vienna, Austria
Email: {john,harald}@ai.univie.ac.at

1 Introduction

Multilingual generation from a single input specification is currently a research topic getting much interest. From a practical point of view it can be used as a possible alternative to machine translation for a range of applications where texts have to be produced simultaneously in different languages. Instead of producing text in a single language and then translating it to the other languages to be covered one can instead automatically generate texts in all the target languages at the same time. One expectation of course is that by using multilingual generation one could avoid many of the inherently difficult problems of fully automatic machine translation.

What becomes clear when looking at the respective system architectures is that the (computer-aided) creation of a monolingual text is replaced by the computer-aided generation of a formal description of the content in some linguistic formalism. The–inherently difficult–parsing task of machine translation becomes obsolete in this scenario.

*The work reported here has been carried out within the LRE Project *GIST* (LRE 062-09) and funded by the Austrian *Forschungsförderungsfonds der Gewerblichen Wirtschaft*, Grant 2/329. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian *Bundesministerium für Wissenschaft, Forschung und Kunst*.

What is clear though is that while in machine translation the depth of representation may vary depending on the type of architecture chosen that for multilingual generation we need some kind of interlingua as a deep linguistic representation. Because, the input specification for multilingual generation must be sufficiently abstract to avoid encoding knowledge in a language specific way–at least with regard to the set of languages considered for the task. It must also be possible to produce specifications for each of the languages with sufficient detail to have the monolingual tactical generation components produce correct text.

In other words, text generation systems producing multilingual output must utilize linguistic knowledge, which has to describe each of the target languages with sufficient accuracy and, in addition, has to perform the mapping of the common input structure to the target text in each of the languages. Thus, knowledge common to all target languages (e.g., the semantics of the target domain) as well as knowledge particular to a single target language (e.g., syntactic and lexical information) has to be represented.

When talking about linguistic knowledge sources supporting multilinguality one must be careful to distinguish between two different ways of usage these sources can support.

- The linguistic formalism[1] is general enough to support the description of the relevant data for a range of different languages, and also may contain all the relevant descriptions for all target languages. However, to make use of these descriptions, for each language a different angle to view them is required. Thus, such knowledge sources are not able to account for different languages simultaneously.

- The linguistic resource is not only able to represent the information needed for all target languages, but also to put it to use simultaneously, making use of the same set of descriptions. Such a linguistic knowledge source may well deserve to be called an interlingua.

This second approach is both more useful but also more demanding. In the following we will show that - not surprisingly - many of the problems

[1]By linguistic formalism we mean a formalism embodying some underlying linguistic theory, i.e. an LFG implementation would count as a linguistic formalism while PATR would not.

faced in MT reappear in this kind of framework although hopefully in a more manageable form.

2 The GIST Setting

The work we will describe in this paper was done in the GIST project. GIST is concerned with the multilingual generation of instructional administrative texts. The target languages of GIST are English, German and Italian. One of the GIST objectives is the reuse of existing resources, in particular the reuse of tactical generators for each of these languages. The system architecture reflects that fact by following a classical division of the generation task into strategic planning and tactical realization. The text planner produces a specification of the text to be generated in ESPL, a language based on SPL [4]. This specification is fed into the three tactical generators which realize the text in the different target languages.

Ideally, this ESPL specification would look the same for each tactical generator. However, this turned out to be impossible. One reason for this fact is that SPL-like languages convey not only semantic information, but also transport syntactic features, which are of course language dependent. Although in the design of ESPL the amount of such syntactic specifications has been kept to a minimum, the requirement of having a separate ESPL statement for each language is a fact.

The semantic part of an ESPL specification, however, should be more or less constant across languages (perhaps modulo different conceptualizations). The basis underlying the description language is a Generalized Upper Model [1], which–together with the domain model–should provide all the linguistic and factual knowledge needed for the generation task. In particular, since the UM is claimed to support multilinguality of the second kind, this knowledge should be suitable for all target languages.

The part of the project which is of interest for the topic of this paper is the creation of descriptions of utterances in these three languages from a single specification and in particular this specification itself.

3 Case study

We will now describe two examples for the kind of problems to be tackled in a realistic environment. In this discussion we assume some basic familiarity with the UM as described in [2].

In particular, we will show cases where the monolingual descriptions cannot be simply merged but where some refinement on the linguistic model must be undertaken. According to [3] we can classify problematic cases into three classes:

- Identity: the distinction can be upheld across the set of languages.
- Extension: If one language is more specific than the others than the set of features and values must be further refined to accommodate for the more specific language.
- Cross-classification: The languages partition the phenomenon in ways which are different to each other. The solution proposed for this case is to find a new semantic description for the phenomenon and to categorize all languages anew.

The first example is concerned with determiners (and in particular articles) for noun phrases. At a first glance the problem may look rather simple. In all three languages involved there are the same three possibilities:

- definite article (*the, der/die/das, il/la*)
- indefinite article (*a, ein/eine, un/una*)
- no article

Looking at a descriptive grammar of any of the three languages one could get the impression that the use and distribution of these different possibilities is analogous. Definite article is used for referring to already known entities, indefinite to introduce new entities, while no article is used for some special cases like proper names and mass nouns which can be marked lexically. Also, since there is no plural form of indefinite articles, the bare form has to be used for this case too.

The first attempt thus was to take this as a case of identity. It was encoded in ESPL using the two features `:identifiability` (values: `identifiable` and `nonidentifiable`) and `:name` were used to discriminate between the three cases. A closer look at the existing corpus of administrative texts reveiled though that the situation is more difficult:

a. Sie haben Anspruch auf Alterspension ab ...
You have the right to a retirement pension from ...

b. Eine Alterspension wird nicht gewährt, wenn ...
A retirement pension is not granted, if ...

This example shows that the set of keywords and values is not sufficient to describe the data from the German corpus. *Alterspension* occurs in *a.* without a determiner, although it is neither a name, a mass noun or plural, as example *b.* shows. Similar cases can be found in the other languages as well. A preliminary investigation showed that the dropping of the article seems to occur occurs in cases of abstract or generic use of common nouns but cannot be derived from grammar or lexicon. Thus such usage has to be specified in the input.

The amendment in this case seems simple: ESPL has to be enriched to provide the corresponding keywords. This amounts to what in [3] is called extension. There are two problems with this solution though:

- The three languages differ somewhat in their attribution of this status to common nouns, this means that the solution cannot be employed if we want to stick to a single representation for all three languages.
- Second, the problem is only shifted one level upwards. Because the component producing the language-specific input must of course be able to decide whether article dropping is appropriate in a certain case or not.

What would really be needed is a deeper understanding of the underlying phenomenon which could lead to a more fine-grained semantic description which can be used in the interlingua. But, of course, one must be aware that this again would make the generation of specifications in this interlingua more complex.

The second example is one which–in contrast to determiners–is known to be a difficult one: the selection of the appropriate temporal preposition. The Generalized UM provides a hierarchy of temporal relations which is claimed to cover the range of all semantically possible temporal relations. Due to the semantic nature of these relations ESPL statements involving them should have identical specifications for all three languages. It turns out, however, that the hierarchy is biased towards the English use of temporal prepositions and does not provide sufficient information for, e.g., German. Consider the examples given in Table 1.

seit	Ich bin seit 2 Wochen auf Urlaub. *I have been on holiday for 2 weeks*
	Ich bin seit 30.8 auf Urlaub. *I have been on holiday since August 30. – I am still on holiday*
	Er war seit 2 Wochen auf Urlaub als der Unfall passierte. *He had been on holiday for 2 weeks when the accident occurred.*
	?Er war seit 30.8. auf Urlaub als der Unfall passierte.
vor	Ich war vor 2 Wochen auf Urlaub. *Two weeks ago I was on holiday. — Now I am back.*
	Ich war vor dem 30.8. auf Urlaub. *I was on holiday prior to August 30.*
	Ich werde vor dem 30.8. auf Urlaub fahren. *I will go on holiday before August 30*
von...bis	Ich fahre vom 20.8 bis 30.8 auf Urlaub. Ich werde vom 20.8. bis 30.8 auf Urlaub fahren. *I'll be on holiday from August 20 to 30.*
	Ich war vom 20. bis 30.August auf Urlaub. *I was on holiday from August 20 to 30.*

Table 1: Some temporal prepositions in German

The preposition *seit* is usually followed by an NP expressing a point in time. Its basic semantics is to give the starting point for some event. Therefore, *seit 30.8.* is a clear-cut case. *seit 2 Wochen* is more subtle. *2 Wochen (two weeks)* denotes a period of time. This doesn't imply though

that the PP itself is a point in time in the former case and a period in the latter. In both cases the semantics is the same, we are talking of a process which has a fixed beginning (in the past) and which is still continuing only that in the second case the point in time is described in terms at the time elapsed between this point and now. It seems then to correspond best to English 'since' (even if it isn't always translated as such), which belongs to the `temporal-locating` subdivision in the UM hierarchy.

The preposition *vor* on the surface behaves similar to *seit*. It is also usually followed by an NP expressing a point in time. Its basic semantics is to locate an event temporally prior to the described point in time. But in contrast to the above, 'vor dem 30.8' (*before 30.8*)' and 'vor 2 Wochen' (*2 weeks ago*) do have a different semantics. 'vor dem 30.8' does not include the 30.8 but refers to the time preceding. 'Vor 2 Wochen' also describes a point in time but in this case the 'vor' is used to convert the time interval '2 Wochen' into a point in time by relating it to now. As a consequence the prepositional phrase as a whole functions rather like a temporal adverb (like 'gestern' *yesterday*). This means it gives an indication about the time the event took place instead of designating a point in time prior to which the event ended. 'Vor' in the former case we can regard as `anterior nonextremal`. 'vor' in the latter certainly belongs in a different place in the hierachy. We suggest `relative non-exhaustive extent`, so long as this would be the correct place in the hierarchy for English 'ago', to which the expression most closely corresponds.

The expression 'von (time1) bis (time2)' (*from ... to*) presents us with a number of problems. 'Bis' is unambiguous in our grid as `anterior-extremal`, but instead of 'ab' or 'seit' which we might expect (in connection with this) we have 'von' which doesn't normally occur as a temporal expression at all. The temporal expressions phrases in the grid below all have either a marked start or finish point, with 'von.. bis' we have both start and finish point given. It is thus like English 'between' which isn't' covered by the UM. (Incidentally, English has exactly the same problem with the expression 'from ..to', 'to' does not normally occur with time expressions. Thus we suggest analyzing 'von' as `ordering posterior` and not attaching it to the extremal/non-extremal division.

Of course, this is only a very preliminary solution. The more general observation must be that temporal relations are expressed in quite different ways in different languages. Information is spread between temporal prepo-

sitions and adverbials and the tense and aspect system of the verbs.

The balance between these means varies widely across languages (E.g., while English has progressive tense, German must express aspect by making use of prepositions). A thorough semantic classification of the whole complex of temporal relations would have to take into account all the various syntactic and semantic means languages have at their disposal.

The problem is that it is not clear what this new ontology should look like. Furthermore, even if such a more fine-grained description could be found we still have to face the problem that this makes the task of generating the descriptions in such an interlingua extremely difficult since in most cases our input descriptions will not be fine-grained enough.

4 Conclusion

In this paper we have tried to show that some of the problems one tries to avoid by replacing machine translation by multilingual generation do pop up again when one takes the notion of multilingual generation serious. In particular, one is again committed to develop an interlingua for the set of languages covered by the generator and this interlingua must be semantically fine-grained enough to support a description abstract enough to accommodate all the languages involved.

However, even with these difficulties in mind, the task for multilingual generation is still more manageable then for machine translation. The reason for this claim is twofold:

1. All the ambiguities arising from parsing a natural language input to arrive at the interlingual representation can be avoided. Multilingual generation starts from a specification in a formal language which may be tailored to processing needs. The parser itself is replaced by the text planner constructing the interlingual representation out of the high level specification. Thus the interlingual representation can safely be expected to be more precise as a one obtained by parsing natural language.

2. The interlingual specification can be made much richer than one obtained by parsing. In particular, features that are hardly derivable from

natural language input given state-of-the-art technology (e.g., pragmatic features) can be specified frankly in the input to the generator, leading to means to guide the generators more precisely.

Thus, even if some MT problems reappear, multilingual generation-where it is applicable-is the better choice for obtaining texts in different languages with the same content.

References

[1] Bateman J., B. Magnini and F. Rinaldi: The Generalized {Italian, German, English} Upper Model, *Proceedings of the ECAI-94 Workshop on Implemented Ontologies*. Amsterdam, 1994.

[2] Henschel R.: Merging the English and the German Upper Model, Technical Report, GMD/Institut für integrierte Publikations- und Informationssysteme, Darmstadt, 1993.

[3] Hovy E. and S. Nirenburg: Approximating an interlingua in a principled way. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, New York, 1992.

[4] Kasper, R. T.: A flexible interface for linking applications to Penman's sentence generator, in *Proceedings of the DARPA Speech and Natural Language Workshop*, Philadelphia, 1989.