Testing Multilinguality in MILE

Jan Odijk

University of Utrecht, UiL-OTS, Trans 10, 3512 JK Utrecht, The Netherlands
jan.odijk@let.uu.nl,
WWW home page: http://let.uil-ots.nl

Abstract. In this report I consider to what extent the multilingual aspects of two different machine translation (MT) systems can be expressed in MILE, the standard proposed in the ISLE project for representing multilingual aspects of lexical entries. The two MT systems are the interlingual system *Rosetta*, and the transfer system *Globalink*. The comparison is based on material available in November and early December 2002. I conclude that, based on this material, it is difficult to evaluate the MILE as a proposed standard properly, and that a detailed demonstration of how the MILE can be applied for these systems is desirable. I hope that the small experiments I carried out on the basis of preliminary and incomplete descriptions of MILE may contribute to a well-defined and usable finalized version.

1 Introduction

In this report I consider to what extent the multilingual aspects of two different machine translation (MT) systems can be expressed in MILE, the standard proposed in the ISLE project for representing multilingual aspects of lexical entries. The two MT systems are the interlingual system *Rosetta*, and the transfer system *Globalink*. One major purpose of setting up standard for multilingual lexicons is to make it possible to exchange multilingual lexicons among language technologies and applications in order to promote maximal reuse of these lexicons. The comparison is based on material available in November and early December 2002. The material available at this time was manyfold. Concerning monolingual aspects, there is a detailed description in the ISLE deliverables (see Lenci (2002)). The monolingual aspects are based on the Simple / Parole / Genelex model (with minor modifications). An implementation by means of an RDF Schema is given but is as yet incomplete. Concerning the multilingual part, several documents contain considerations, questions, architectural design issues, design requirements and desiderata, strategy, recommendations, and methodology (Grishman (2001), Bertagna *et al.* (2000), Calzolari *et al.* (2001a)). Other documents specify requirements from existing multilingual dictionaries, models, and MT transfer dictionaries (Villegas (2001), Calzolari *et al.* (2001b), Thurmair (2001), Takenobu (2001), Zajac (2001), Odijk (2000). Yet another document describes a tool that, given a DTD, operates on SGML documents to create a relational database for multilingual lexicons and a user interface to it (Bel and

Villegas (2001)) However, at this point in the project, no concrete proposal, or a consolidated set of guidelines for multilingual aspects existed yet. I conclude that, based on the material available at this point in the project, it is difficult to evaluate the MILE as a proposed standard properly, and that a detailed demonstration of how the MILE can be applied for these systems is desirable. I hope that the small experiments I carried out on the basis of preliminary and incomplete descriptions of MILE may contribute to a well-defined and usable finalized version.

This report is organized as follows. Section 2 deals with the Rosetta MT system: it is introduced and briefly described in section 2.1, and a comparison with MILE is carried out in section 2.2. Section 3 deals with the Globalink system. It is briefly described in section 3.1, and an attempt to compare it to MILE is carried out in section 3.2. Section 4 summarizes the main conclusions.

2 The Rosetta MT system

In this section I very briefly describe some of the crucial properties of the approach adopted in the Rosetta MT system that are relevant to the topic of this report, and compare them to the proposed MILE standard.

2.1 Introduction to the Rosetta MT system

The Rosetta MT system is a research prototype translating between Dutch, English and Spanish. It was developed in the Rosetta project carried out at the Philips Research Laboratories. For an extensive description of the approach and the system developed, I refer to Rosetta (1994).

The grammars of the Rosetta system are a special kind of *compositional* grammars called *M-grammars*. A compositional grammar consists of *basic expressions* and *rules*. One can derive utterances by recursively applying rules, initially to basic expressions. In order to be able to deal adequately with the complexities of natural language, rules apply to (one or multiple) syntactic trees called *S-trees* (short for surface trees) that encode syntactic categories, attribute-value pairs, linear order, and constituent structure. Basic expressions are a special kind of S-tree called a *lexical S-tree*. The way an utterance is derived can be represented by recording which rules must be applied, and to which arguments, initially basic expressions, in a syntactic derivation tree or *syntactic D-tree*.

Compositional grammars are designed in such a way that the principle of *Compositionality of Meaning* holds. This principle can be stated as follows:

(1) **Compositionality of Meaning**
The meaning of an expression is a function of the meaning of its parts and the way they are combined

This is achieved in compositional grammars because both basic expressions and rules have a meaning. Therefore, the meaning of an utterance can be derived in parallel with the syntactic derivation, and the way the meaning is computed

can be represented in a *semantic D-tree* consisting of unique identifiers for the meanings of the rules and the meanings of the basic expressions.

The principle behind the method used to translate with M-grammars is called the principle of *Compositionality of Translation*:

(2) **Compositionality of Translation**
Two expressions are each other's translation if they are built up from parts which are each other's translation, by means of rules which are each others translation

Having the same meaning is a necessary condition for translational equivalence, but factors additional to meaning (e.g. stylistic) might be relevant in determining translational equivalence.

Compositionality of Translation is achieved by attuning the M-grammars G1 and G2 of two languages, as follows:

- For each basic expression in grammar G1 there must be at least one basic expression in G2 that is translationally equivalent
- For each rule in grammar G1 there must be at least one rule in G2 that is translationally equivalent

Such attuned grammars are called *isomorphic grammars*. For each derivation tree in G1 there will be an isomorphic derivation tree in G2 that is translationally equivalent, and if the grammars satisfy certain additional restrictions, one can derive translationally equivalent utterances from G1 and G2 in parallel.

I have discussed the basic properties of the method of compositional translation only by speaking of the grammars (i.e. abstract characterizations of the language) used in this method. M-grammars satisfy a number of additional conditions that make them suitable to be used in actual analysis and generation of utterances. These, however, are less relevant for the topic of this report, and I will not discuss them here. For details, I refer to Rosetta (1994).

The method of compositional translation, when applied to multiple languages, naturally leads to an interlingual system. However, it is a very special kind of interlingual system. First, the interlingua in this approach is not an independently defined or existing language, but rather results automatically as a consequence of attuning the grammars: the interlingua consists of the semantic derivation trees, which express translational equivalence: the semantic derivation trees are shared by all languages for which attuned grammars have been created. Second, the interlingua is defined and valid only for the languages for which attuned compositional grammars have been defined. Third, the method of compositional translation guarantees that the interlingua contains exactly the semantic derivation trees that are defined in the grammars of the languages for which the interlingua holds. Fourth, though the approach naturally leads to an interlingual approach, there is no absolute requirement to adopt an interlingual approach: one can also devise it as a transfer system, or, what is more interesting, one can have a mixed interlingual/transfer system: e.g., interlingual for rules, but transfer for lexical items.

If one adopts a transfer approach, or a mixed interlingual/transfer approach, the very design of the method of compositional translation restricts the kind of transfer allowed in ways that are generally desired: the approach allows for transfer in the sense that the mapping from one language to another is language-pair dependent, and dependent on the translation direction. However, only *simple transfer* is allowed, i.e. only mappings from a single basic expression to a single basic expression, or from a single rule to a single rule. This excludes

- complex transfer, e.g. one or more basic expressions to multiple basic expressions, or vice versa; one or more rules to multiple rules, or vice versa; combinations of rules and basic expressions to combinations of rules and basic expressions, etc.
- conditional transfer: simple transfer that is valid only if certain structural conditions are satisfied
- actions on the target of transfer (e.g. modifying it, adding something to it, etc.)

The choice between an interlingual, transfer or mixed approach can be made on the basis of the intended applications, on the basis of available resources, or perhaps even other factors. The first can be illustrated as follows: if one does not know in advance which target language an utterance will be translated into, the interlingual approach is called for. If one does know the target language in advance, the transfer approach might be more natural and efficient. The second aspect relates to the available lexicons: almost all existing large size translation lexicons are language-pair and translation-direction dependent, and there are hardly interlingua-based translation lexicons of a significant size. So this may force one to deploy the transfer approach, at least for basic expressions.

The lexicons used in the Rosetta3 system are based upon the *Van Dale* contemporary translation lexicons.[1]

A typical example of a lexical entry (for the transitive verb *innemen* (lit. 'to take in') from the Dutch lexicon is given in (3):

(3) neem
: $s_aV_00_inneem
: $m_aV_0001_inneem m1 mbt. geneesmiddelen swallow,
$m_aV_0002_inneem m2 mbt. plaatsruimte take up,
$m_aV_0003_inneem m3 veroveren take,
$m_aV_0004_inneem m4 binnenhalen bring in,
$m_aV_0005_inneem m5 aan boord nemen take on,
$m_aV_0006_inneem m6 inkorten (korter) take up,
$m_aV_0007_inneem m7 inkorten (nauwer) take in,
$m_aV_0008_inneem m8 verzamelen collect
{ monolingual morphological and syntactic properties }

[1] For Dutch and English: Martin (1986) and Martin (1984).

The first line in this example contains the stem *neem* of the Dutch verb *innemen*.[2] The second line gives a unique identifier for the lexical item, taking into account morpho-syntactic aspects only. The identifier ($s_aV_00_inneem for this verb) is called the *syntactic key* or briefly *skey*. The following lines (except the last one) contain meanings and their associated properties of this lexical item. Each meaning has a unique identifier (e.g. $m_aV_0001_inneem for the first meaning given) called the *meaning key* or briefly *mkey*. The various meanings are also ordered by priority. This is indicated by the labels *m1* through *m8*: the lower the number behind the *m*, the higher the meaning priority is. This ordering is used to order the translations when multiple translations are generated, and to select the most plausible one when one translation is needed. Each meaning also contains a disambiguator, i.e. a string in natural language (in the source language) that can be used to allow a (human) user of the MT system to resolve ambiguities that cannot be solved by the MT system itself.[3] This is useful in applications where an author submits his/her text for automatic translation, but where the target language is not known yet at the moment this text is submitted. All the ambiguities that remain after the system has analyzed an utterance must be resolved in the source language by the author of the text, after which the utterance can be stored in the form of its semantic derivation tree. The system will present the disambiguators to the author, so that he/she can select the intended meaning. The last column of the meanings in (3) contains the English translation for each meaning. This is not part of the lexical entry, but has been added here for convenience and clarification. The final line simply states that the morphosyntactic properties of the lexical item (e.g. that it is a transitive verb, that it takes the separable prefix *in*, etc.) are listed here (but the properties themselves, which are identical in nature to the properties of single word lexical items, have not been included here).

2.2 Rosetta and MILE

In the preceding section we have described some of the essential properties of the Rosetta MT system, and the representation of lexical items in such a system. By the very design of the system, the multilingual part has been reduced to the

[2] The verb *innemen* contains the separable prefix *in*, which is not part of the stem, because the verb part and the separable prefix can occur separately as in *Hij* ***nam*** *de pillen* ***in*** 'He swallowed the pills'. A grammatical attribute specifies that this stem must be combined with the separable prefix *in* to count as a form of this lexical item.

[3] The ordering and the meaning descriptions given here have been taken directly from the Van Dale Dutch-English translation dictionary (Martin (1986:618)). Only the expression *iemand voor/tegen zich innemen* 'to make oneself sympathetic/unsympathetic with someone' is left out here, since it is not a separate meaning (rather an idiomatic expression containing the verb) though (incorrectly) represented as such (as the 4th meaning) in the Van Dale dictionary. In addition, the 6th and 7th meaning have been explicitly distinguished here, though they are two different translations of the same meaning in the Van Dale dictionary.

bare minimum, so it should be easy to represent this by means of the MILE standard.

We have seen that a Rosetta lexical item has a number of properties. We start with properties related to monolingual aspects: These include:

- a syntactic key, or skey
- a meaning key, or mkey
- a priority ordering for different meanings
- a disambiguator, intended for use by humans

Can we represent these items in MILE? What do they correspond to?

The *skey* appears to have a direct correlate in MILE, viz. the *SynU id* attribute (see Lenci (2002), section 6.3.3.1).

The *mkey* also appears to have a direct correlate in MILE, viz. the *SemU id* attribute (see Lenci (2002), section 6.3.4.1)

A correlate for priority ordering, however, does not appear to have been defined. Perhaps it may be treated as an instance of SemFeature (see Lenci (2002), section 6.3.4.4), but the property is so basic to a translation lexicon that it deserves a separate treatment of it own in the core of the MILE definition.

There also appears not to be a correlate in MILE for the Rosetta disambiguators. Though important and interesting work on disambiguators has been carried out in the ISLE project (see Atkins and Bouillon (2002)), this does not relate to the problem at hand here: the disambiguators here are intentionally not formalized, because they are intended to be used by humans. There is one candidate attribute that might perhaps function as disambiguator, viz. the *comment* attribute (see Lenci (2002), section 6.3.4.1), but that attribute is too general, and disambiguators are so basic that they deserve a treatment as independent attributes in the core of the MILE definition.

Turning now to multilingual aspects, the very design of Rosetta ensures that nothing special has to be done for the multilingual aspects: tuning the grammars automatically takes care of translational aspects. However, in the Rosetta system, tuning grammars has as a consequence that two identical mkeys from different tuned grammars represent the same meaning, or, more precisely, translational equivalence. Though the documentation is not fully unambiguous, nothing indicates that identity of *SemU id*s from different MILE lexicons has any significance at all: it certainly does not imply sameness of meaning or translational equivalence.[4] As a consequence it is unclear, given the documents I base myself on, how sameness of meaning in an interlingual approach can be represented in MILE, and how the interlingual nature of the system as a whole can be represented.

[4] As confirmed to me by Alessandro Lenci, p.c.

3 The Globalink MT system

In this section I will investigate to what extent the approach adopted in the Globalink system[5] can be represented in the MILE framework. There is nothing particularly special about the Globalink system and its treatment of multilingual aspects. I have selected it as a test example because it is representative of a large class of (commercially available) MT systems (see Thurmair (2001)). I will first introduce some basic properties of the Globalink system and lexicon, and then attempt to compare it to MILE.

3.1 Introduction to the Globalink MT system

The Globalink MT system is a transfer-based MT system. A lexicon in Globalink is represented in a special format called the Lexicon Interchange Format (LIF). This is a textual (comma-separated) representation of lexicons that abstracts from the engine-internal representation. A Globalink lexicon is basically a flat table (a set of records) where each record represents one translation relation for a headword. There is no explicit distinction between source language, target language, and transfer lexicon. Globalink lexicons in LIF-format are bilingual, unidirectional, language-pair and domain-dependent lexicons where each row in the table represents one translation of a source language head word. Monolingual properties are present to the extent that they are necessary for translation.

Fields in a LIF record include

- lemma
- lemma's translation
- lemma's PoS
- morphological properties
- attributes relevant to this translation of the headword (these attributes induce tests on the lemma's context and actions on the target word)
- administrative information

Examples of attributes encoding tests on the lemma's context are, inter alia, properties of syntactic selection (transitivity, intransitivity), semantic type (*Animate, Human, Place,* etc.), and syntactic subcategories (Proper v. common noun etc.).

Attributes that induce actions on the target language word include, inter alia, specifications that grammatical features such as *number* or *PoS* of the target word differ from the corresponding attribute of the source language lemma.

[5] The Globalink MT technology is owned by Bowne. I base the description given here on Odijk (2000), which supplies a description of the properties of Globalink lexicons that was made in the context of the ISLE project when this technology was owned by Lernout & Hauspie Speech Products.

3.2 Globalink and MILE

Can we represent lexicons in LIF format in the MILE format, and how should the mapping be defined? The issue is not simple: LIF makes no explicit distinction between source, target and transfer lexicon, but this is a crucial ingredient of MILE. So a mapping of LIF into MILE would require separating the source language specific properties from the target language specific properties and from the transfer related properties. It also would require separating syntactic from semantic attributes, creating new *synUs* and *semUs*, and assigning the syntactic properties to the newly created *synUs* and the semantic properties to the newly created *semUs*. The newly created synUs and semUs also have to be properly linked to each other, so that the syntactic and semantic properties remain properly associated. Furthermore, since each LIF record expresses one translation of a head word, one LIF record may result in multiple identical synUs for the same word whether in the source language or in the target language. Such identical synUs should be collapsed into a single synU. For the target language aspects, a synU must be made. The relation between source and target language must be made by directly specifying a link between a source language synU and a target language synU, skipping the semantic level.

From this brief sketch it is quite clear that mapping Globalink LIF entries into MILE is perhaps possible, but certainly quite difficult. If two different people would carry out the task, it is unlikely that they would end up with the same representation in MILE — which would make the representations less useful for exchange between different systems. For this reason it would have been useful if the MILE description would have been much more specific on how to map transfer-based systems such as Globalink (and most commercially available MT systems, e.g. Comprendium, Logos, Systran, see Thurmair (2001:3-4)) into MILE and vice-versa.

4 Conclusions

I have considered the proposed MILE standard for multilingual and translation lexicons by attempting to apply it to two completely differently organized MT systems: the Rosetta interlingual system, and the Globalink direct transfer system. I concluded that in both cases applying the MILE was not obvious, and it appears to lack certain essential properties that deserve a position in the core. In particular, based on the information available, it is not clear that the proposed standard can really serve the purpose of making exchanges of multilingual lexicons between completely different NLP systems possible.

The comparison was carried out on the basis of preliminary and incomplete descriptions of MILE. I hope that the exercise I carried out with these incomplete descriptions of MILE may contribute to a well-defined and usable finalized version, that tackles the issues I brought up in this small report, so that a standard results that can effectively increase the reuse of multilingual lexicons across different MT (and other NLP) systems and applications.

References

[2002] Sue Atkins and Pierrette Bouillon. Sense indicators in the lexicographic process - lexicographic sense indicators as candidates for transfer conditions - the lexicographic approach: Clues in bilingual dictionaries and the DB - with demo. ISLE Presentation, Pisa 2 december 2002, December 2002.

[2001] Nuria Bel and Marta Villegas. MILE tool documentation: Preliminary report. ISLE Report WP4, Gilcub, Barcelona, February 2001.

[2000] Francesca Bertagna, Nicoletta Calzolari, Alessandro Lenci, and Antonio Zampolli. The multilingual ISLE lexical entry: A discussion paper. Pisa, December 2000.

[2001a] Nicoletta Calzolari, Ornella Corazzari, and Sergio Rossi. Recommendation of best practices / guidelines / standards. ISLE Computational Lexicon Working Group, March 2001.

[2001b] Nicoletta Calzolari, Ralph Grishman, and Martha Palmer. Survey of major approaches towards bilingual/multilingual lexicons. ISLE deliverables 2.1-3.1, February 2001.

[2001] Ralph Grishman. Building a MILE entry. Presentation, March 2001.

[2002] Alessandro Lenci. Standards and best practice for multilingual computational lexicons: MILE (the multilingual ISLE lexical entry). First partial draft of deliverables 2.2 & 2.3. ISLE deliverables, Pisa, November 2002.

[1984] Willy Martin, editor. *Groot Woordenboek Engels-Nederlands*. Van Dale Lexicografie, Utrecht, 1984.

[1986] Willy Martin, editor. *Groot Woordenboek Nederlands-Engels*. Van Dale Lexicografie, Utrecht, 1986.

[2000] Jan Odijk. Multilingual lexicons. ISLE Meeting, Paris, April 2000.

[1994] M.T. Rosetta. *Compositional Translation*, volume 273 of *Kluwer International Series in Engineering and Computer Science (Natural Language Processing and Machine Translation)*. Kluwer Academic Publishers, Dordrecht, 1994.

[2001] Tokunaga Takenobu. Towards the multilingual ISLE lexical entry: Some examples from Japanese. ISLE presentation, 2001.

[2001] Gregor Thurmair. Multilingual dictionaries: Transfer part. ISLE Memo, March 2001.

[2001] Marta Villegas. Multiword expressions in Genelex model. March 2001.

[2001] Rémi Zajac. Translation and semantics of MWEs from a machine translation point of view. March 2001.