Shake-and-Bake MT and Morphology*

Davide Turcato
turk@cogsci.ed.ac.uk

1 Introduction

We will address the issue of designing MT systems where the capability to reuse grammars taken off the shelf is maximized, thus restricting the need for application-specific work. Or. to put it in a slightly different way, our proposal is aimed at encouraging the development of pure grammars (i.e. grammars describing languages in the abstract), to be contrasted with those which are extensively tailored for a specific translational use.

Our proposal is set on the background of the Shake-and-Bake approach to MT (Whitelock, 1991). In this model translation is achieved by putting in correspondence the bag of lexical items employed in a source sentence with a corresponding bag of lexical items in the target language. The translation of the source sentence is whatever target sentence can be built out of the lexical items in the target bag by freely permuting its elements. We will propose an extension to a standard Shake-and-Bake system in order to support a wider reusability between applications and to give grammar writers more freedom. We will focus on the morphological component of a grammar, which seems to be one of the main bottle-necks which restricts a full portability of grammars under a Shake-and-Bake MT system. It will be shown that the Shake-and-Bake system architecture outranks a number of morphological approaches available in the linguistic literature and it will be discussed how such a restriction can be removed.

2 The issue of morphology

Assuming without further discussion that the Shake-and-Bake approach successfully supports portability to new algorithms, directions and language pairs, we will focus on the following issue: what kind of restrictions the Shake-and-Bake architecture places on the grammars to be used in the system? In other words, at what extent grammars falling in the broad domain of declarative, lexically-based grammars can be accommodated in a Shake-and-Bake MT system? In answering the question we will focus on the morphological component of grammars.

* I would like to thank Dr. Chris Brew (HCRC, Edinburgh) who helped me greatly in doing the research that led to this paper.

2.1 Some approaches to morphology

In first place, we will give an account of some approaches to morphology available in linguistic theory. Our overview doesn't claim by any means to be complete. It is only meant to show a sample of different morphological models, which we will refer to in the subsequent discussion.

Morpheme-based morphology. In morpheme-based morphology each grammatical and inflectional formative, in addition to lexical stems, is represented by means of a specific morpheme, i.e. a lexical sign. Thus, inflected forms and fully specified lexical signs are formed by combining lexical stems with formatives, by means of word formation rules of the same kind as phrase structure rules.

HPSG morphology. In HPSG the internal structure of words is accounted for by means of two devices: *a multiple inheritance hierarchy and lexical (redundancy) rules.* Each node in the hierarchical structure is a lexical entry, but only the leaves of the hierarchy are *actual* lexical entries. The intermediate nodes are *generic* lexical entries, i.e. underspecified entries which contribute to the formation of subordinate actual entries. Grammatical formatives fall in the latter class. Lexical rules avoid redundant information in the lexicon, allowing generalizations of the inflectional kind.

Realizational morphology. In realizational morphology, according to Anderson's version (1986). morphemes are replaced by rules. A rule takes as input a pair of a lexical stem and a complex symbol representing the inflectional categories associated to the former. Anderson (1986) provides the following general form for an inflectional rule:

$$(S,M) \Rightarrow (S', M')$$

where S is the item and M is its morpho-lexical representation. The rule schema shows that both of them can be modified. The item can undergo several kind of changes other than just concatenation with some affix. Thus, whatever contribution is given to a stem by its combination with a morpheme in a derivational system, in the present approach is obtained by submitting the stem and its representation to a rule. As pointed out by Erjavec (1994, p. 1), in the realizational approach ". . . it is argued that the minimal phonological unit over which morphological generalizations can be made consistently is the whole word, or at least its stem, and that concatenation is only a special case of phonological realization".

2.2 Morphology in a Shake-and-Bake system

In the Shake-and-Bake model "the treatment of grammatical equivalences assumes a sign-based morphology in which inflections and other grammatical items are simply lexical signs". (Whitelock, 1991, p. 229). The committment to a morpheme-based morphology is not dictated by any linguistic considerations, but rather required by the architecture of the system. Whatever in a sentence is relevant to the translation process must be represented as a proper lexical sign, if it is to be used in the bilingual lexicon. As far as grammatical and inflectional formatives are concerned, some simple examples should be enough to make clear that they are relevant elements in stating translational equivalence, particularly when highly inflectional languages are involved. The well known phenomenon of head-switching, which we restate here for the English-Italian language pair, is one such example:

(1) John runs up the street.
(2) Giovanni sale la strada di corsa.

In the example above, the following basic equivalences are established, among the others:

(3) {run} ≡ {di, corsa}
{pres} ≡ {pres}
{up} ≡ {salire}

It would be problematic to state the equivalence relation if a formative pres were not there, since the present tense is related in the two sentences to the non-equivalents items *run* and *salire.*

Let's consider another example.

(4) a. I walk. [lo] cammino.
b. I will walk. [lo] camminerò.
c. I would walk. [lo] camminerei.

In this case the most elegant and meaningful way to state basic equivalences is the following:

(5) {walk} ≡ {camminare}
{will} ≡ {future}
{would}≡ {conditional}

according to which the tensed Italian forms result from the combination of a base form with a tense formatives. Such examples should provide enough evidence that the linguistic aspects represented by formatives are relevant in translation equivalences and thus it would be desirable to have formatives available in the bags of items on which equivalence is stated.

2.3 The problem

Although the morpheme-based morphology fits well in the Shake-and-Bake model and provides a satisfactory treatment of morphological aspects, it seems that competing morphological models cannot be accommodated as easily. In the HPSG approach, for instance, formatives are generic lexical entries, which never appear in bags of lexical items underlying sentences. In realizational morphology they don't even appear as generic lexical entries. In both cases they can only be found as feature bundles embodied in some larger sign. In HPSG, for instance, the Italian verbal form *camminerò* would be a basic morpheme partially specified as follows:

```
(6) word
    synsem:loc:cat:head:verb
    synsem:loc:cat:head:vform:future
```

At no level an actual lexical sign would represent the *future* formative. Thus, it would be problematic to state the translational equivalences of the examples above, like that between the *future* Italian formative and the English form *will.*

To sum up, the Shake-and-Bake architecture places severe and linguistically unmotivated restrictions on the range of grammars which can be used in the system. The domain of declarative, lexically-based grammars is narrowed down to those which adopt a morpheme-based morphology. In the next section we will address the task of removing such a restriction.

3 Extension of a Shake-and-Bake MT system

3.1 Outline

In the original Shake-and-Bake system a unique lexicon works as an inventory of signs which can be used in parsing and an inventory of signs which can be used in translational equivalences. However, under different approaches to morphology, the lexicon used in parsing is no longer adequate to be an inventory of possible signs for translational equivalences. To make clear the double use of the lexicon just described, we will introduce a terminological distinction, to which we will refer in the following. We will refer to the lexicon in the former sense as a *g-lexicon* ('g' for 'grammar') and to the latter as a *t-lexicon* ('t' for 'translation').

While under a derivational approach g-lexicon and t-lexicon are identical, what is required in order to allow a different morphological approach is a relaxation of such an identity: g-lexicon and t-lexicon should be allowed to be different sets. Thus, our proposal, to state it in the most general way, is to introduce in the MT system a mapping between a g-lexicon and a t-lexicon, such that, for every sign in the former, the latter contains a corresponding suitable bag of signs for translational purposes. We will refer to it as a *lexical mapping.*

Roughly, what this mapping should do is extracting from a lexical sign those substructures which could possibly undergo the equivalence relation as autonomous items. Arguments for which the mapping is not otherwise defined are mapped onto themselves. The mapping should be reversible. Although what we have said so far referred implicitly to signs of a source language, the same mapping should be used to map signs of a t-lexicon to signs of the g-lexicon of some target language, there to be used for generation.

3.2 The structure of the lexical mapping

We define a lexical mapping as a set of rules, which we name *splitting rules.* Each rule is an ordered pair < *Condition, Formative* > where both the elements are feature structures. The former contains the specification of a restriction which the lexical sign must satisfy in order to undergo the mapping, the second specifies the feature structure which has to be 'extracted'. Thus, a feature structure *X* matches a splitting rule if *X* subsumes both the *Condition* and the *Formative* specified in the rule. To summarize:

(7) a *lexical mapping* exists between a g-lexical sign *X* and a pair of t-lexical signs < T_1, T_2 > if there is is a splitting rule S = < C, T_2 > such that:

1. *C* subsumes *X;*
2. T_2 subsumes *X;*
3. T_1 is the complement of T_2 in *X.*

The definition can be straightforwardly generalized in order to deal with bags of signs. The rest of the section will be devoted to providing a clear and formally precise statement of the notion of complementation between feature structures, which is the core notion in order to clarify the substructure relation and the extraction operation.

3.3 Complementation on feature structures

Untyped feature structures.

(8) Let *X* and *Y* be feature structures. A feature structure *Z* is a *complement* of *X* in *Y* if and only if:

1. *X* subsumes *Y;*
2. for every atomic-valued path *P* in *Y:*
   (a) if *P* shares its value with a path *P'* and exactly one of the two paths is defined for *X,* then *P* is defined for *Z*, with identical value; otherwise:
   (b) i. if *P* is defined for *X,* then *P* is not defined for *Z*;
   ii. if *P* is not defined for *X*, then *P* is defined for Z, with identical value;
3. no other atomic-valued path is in *Z.*

Typed feature structures. A system of totally well-typed feature structures will be assumed.

(9) Let *X* and *Y* be typed feature structures. A typed feature structure *Z* is a *complement* of *X* in *Y* if and only if:

1. *X* subsumes *Y;*
2. for every path *P* in *Y:*
   (a) if *P* is defined for *X*, then:
      i. *P* is not defined for *Z* or
      ii. *Type(Z,P)* subsumes *Type(X,P);*
   (b) if *P* is not defined for *X*, then:
      i. *P* is defined for *Z* and
      ii. *Type(Z,P) = Type(Y,P);*
3. no other path is in *Z.*
4. there is no typed feature structure which satisfies the conditions above and subsumes *Z.*

The last condition is added in order to ensure minimality of the resulting feature structure. If that condition had not been added, several feature structures would always satisfy the definition. Among them, the input feature structure *Y* would always trivially do.

We still have to take into account path-sharing. Rather than restating formally the above definition we will provide a ‘procedural’ account. The problem is that two paths sharing their value could receive different values in virtue of the definition above. In order to avoid this sort of mismatch we add the further requirement that two paths sharing their value in *Y* and both present in *Z* must share their value also in the latter. Such a value must be identical to the more general of the two values which the coindexed paths receive in virtue of the definition above.

4 Conclusion

The introduction of a lexical mapping and the distinction between a lexicon used for monolingual purposes and a corresponding lexicon used in translation sets free monolingual components from restrictions due to the architecture of the MT system. A wider range of grammars can be used, since the result previously guaranteed by the assumption of a morpheme-based morphology is provided by the lexical mapping.

Let’s consider, for instance, the example in (4). A lexical mapping could contain a splitting rule like the following:

(10) `<synsem:loc:cat:head:verb, synsem:loc:cat:head:vform>`

which says that, whenever a verbal form is in a g-lexical bag, its aspect must be in the corresponding t-lexical bag as an independent item. Thus, given (11) as a (partial) sign for an Italian verb *camminerò,* (12a) and (12b) would appear in the t-lexical bags, to be put in correspondence, respectively, with the English signs for *walk* and *will.*

(11)
```
word
synsem:loc:cat:head:verb
synsem:loc:cat:head:vform:future
```

(12a)
```
word
synsem:loc:cat:head:verb
```

(12b)
```
synsem:loc:cat:head:vform:future
```

It is worth noting that the system proposed here properly extends the original system. If a morpheme-based approach is in use. the lexical mapping is trivially defined as an identity mapping. Thus, the original system can be transferred under an extended system without any changes.

References

[1] Anderson, S.R. (1986) Disjunctive ordering in inflectional morphology. *Natural Language and Linguistic Theory.* 4 (1), pp. 1-32.

[2] Erjavec, T. (1994) Realizational Morphology in ALE. Technical report, Integrated Language Database Project Deliverable.

[3] Pollard, C. and I.A. Sag (1987) *Information-Based Syntax and Semantics,* Vol. 1: Fundamentals. Stanford, Ca.: Center for the Study of language and Information.

[4] Pollard, C. and I.A. Sag (1994) *Head-Driven Phrase Structure Grammar.* CSLI and University of Chicago Press. Stanford, Ca. and Chicago, 111.

[5] Whitelock, P. (1991) Shake-and-Bake Translation. In M. Rosner, C.J. Rupp and R. Johnson, eds., *Draft Proceedings of the. workshop held 18-20th September 1991,* pp. 215-239, Istituto Dalle Molle IDSIA.