Translation Equivalence and Lexicalization in the ACQUILEX LKB

Antonio Sanfilippo, Ted Briscoe, Ann Copestake
Computer Laboratory, University of Cambridge
New Museum Site, Pembroke Street
Cambridge CB2 3QG, UK
Antonio.Sanfilippo@cl.cam.ac.uk, Ted.Briscoe@cl.cam.ac.uk
Ann.Copestake@cl.cam.ac.uk

Maria Antonia Marti, Mariona Taule
Departamanet Filologia Romanica
Secc. Linguistica General
Av. Gran Via Corts Catalanes 585
08007- Barcelona, Spain
fibcls04@lsi.upc.es

Antonietta Alonge
Istituto di Linguistica Computazionale
CNR
Via della Faggiola 32
56100 Pisa, Italy
leminter@icnucevm.cnuce.cnr.it

Abstract

We propose a strongly lexicalist treatment of translation equivalence where mismatches due to diverging lexicalization patterns are dealt with by means of translation links which capture cross-linguistic generalizations across sets of semantically related lexical items. We show how this treatment can be developed within a unification-based, multilingual lexical knowledge base which is integrated with facilities for semi-automatic development of bilingual lexicons, and describe an approach to machine translation where generation difficulties arising from the lexicalist approach to complex transfer can be solved without making special assumptions about phrasal transfer.

1 Introduction

The treatment of translation equivalence is probably one of the most difficult questions to grapple with in setting up a lexical component for MT systems (Tsujii, 1991). Languages are known to exhibit distinct preferences in lexicalization patterns (Talmy, 1985) in such a way that translation may give rise to complex lexical and structural relations. However, since complex translation equivalences often involve classes of lexical items at a time, the use of idiosyncratic transfer rules which operate on instantiated phrasal analyses is linguistically unmotivated and undesirable from an engineering perspective, as they will be time-consuming to construct.

Consider the case of movement verbs. In English, it is usually possible to integrate movement verbs which specify the manner in which motion occurs (e.g. *swim, walk, stagger, crawl)* with a locative expression of completed path, e.g. *across the river, to the pub;* by contrast, such a possibility is seldom available in languages such as Spanish or Italian (Talmy, 1985; Jackendoff, 1990). For example, an acceptable translation for the English sentence *John swam across the river* in Spanish would be *Juan cruzó el río nadando* where English *swim across NP* is translationally equivalent to *cruzar NP nadando* ('cross NP swimming'; Beaven, 1992).

Translation mismatches of this complexity present difficulties with respect to both lexical representation and generation. Because the mismatch in this case arises from diverging regimes of lexicalization,

it would be desirable to state the equivalence at the lexical level; this can be done using lexical transfer techniques, e.g. *bilingual signs* (Beaven & Whitelock 1988; Whitelock, 1988; Zajac, 1989; Estival *et al.,* 1990; Tsujii, 1991). At the same time, one side of the equivalence (the Spanish side in our example) involves a phrase with a 'gap' inside (i.e. the 'goal' argument) which can only be filled after the input source-language string is analyzed. For generation purposes, it would thus be more convenient to establish the translation equivalence through structural correspondences which relate phrasal descriptions in the source and target languages, following the treatment of *head switching* mismatches proposed by Kaplan *et al.* (1989). However, the use of phrasal transfer to cope with lexically governed mismatches requires the creation of specialised equivalents of phrase structure rules restricted to specific lexical semantic classes which are unmotivated from the perspective of the monolingual grammars.

The goal of this paper is to present a strongly lexicalist approach to translation equivalence where lexical transfer can be made to drive generation without direct reference to phrasal transfer.

2 Background

Within the ACQUILEX project,[1] we are testing the feasibility of constructing a multilingual Lexical Knowledge Base (LKB) for a large subset of nouns and verbs using monolingual lexicons semi-automatically derived from English, Spanish, Dutch and Italian machine-readable dictionaries (Copestake, 1992; Sanfilippo & Poznański, 1992; Ageno *et al,* 1992; Vossen, 1992; Calzolari, 1991). The ACQUILEX LKB provides a lexicon development environment which uses a typed graph-based unification formalism as representation language.[2] It allows the user to define an inheritance network of types with associated features, and to create lexicons where such types are semi-automatically assigned to lexical templates which encode word-sense specific information extracted from machine-readable dictionaries. Consider, for example, the LKB entry relative to the first sense of the verb *swim* in the *Longman Dictionary of Contemporary English* (Procter, 1978) where **STRICT-INTRANS-SIGN** provides a general characterization of (strict) intransitive verbs, the *psort* **DEFAULTS-STRICT-INTRANS-SIGN** introduces default values, and the remaining types on the right side of path equations express syntactic and semantic properties more specific to the word sense under analysis including dictionary information (sense-id).[3]

(1) swim L_1_1

```
STRICT-INTRANS-SIGN
<> < DEFAULTS-STRICT-INTRANS-SIGN <>
< cat : result : m-feats : reg-morph > = FALSE
< sen : ind > = PROC
< cat : active : sem : arg2 > = (E-ANIMAL E-HUMAN)
< cat : active : sem : pred > = P-AGT-CAUSE-MOVE-MANNER
< sense-id : dictionary > = "LDOCE"
< sense-id : ldb-entry-no > = "36080"
< sense-id : sense-no > = "1"
< sense-id : sem-field : set-header > = "particular-ways-of-moving"
< sense-id : sem-field : set-group > = "Moving-coming-and-going"
< sense-id : sem-field : set-main > = "Movement-location-travel-and-transport".
```

When the LKB entry above is loaded, all the constraints associated with its type specifications are expanded giving rise to the feature structure representation in Figure 1.

The construction of bilingual lexicons is carried out by establishing translation links (henceforth *tlinks)* between LKB entries which represent word senses from distinct monolingual dictionaries. The assignment of *tlinks* to LKB entries is semi-automatically driven by statistical comparison of feature structure representations of word senses (Copestake *et al.,* 1992). For our purposes, it will suffice to

[1] *The Acquisition of Lexical Knowledge from Machine Readable Dictionaries,* ESPRIT BRA 3030.

[2] A detailed description of the LKB's representation language is given in papers by Copestake, de Paiva and Sanfilippo in Briscoe *et al.* (forthcoming); various properties of the system are also discussed in Briscoe (1991), Copestake (1992) and Sanfilippo & Poznański (1992).

[3] The information relative to the attribute sem-field concerns the semantic codes of the *Longman Lexicon of Contemporary English* (McArthur, 1981) which were used in the individuation of semantic verb classes (Sanfilippo & Poznański, 1992).

point out that the key element of this approach to constructing multilingual lexicons semi-automatically is the use of a common type system to encode syntactic and semantic properties of lexical items in the four languages.

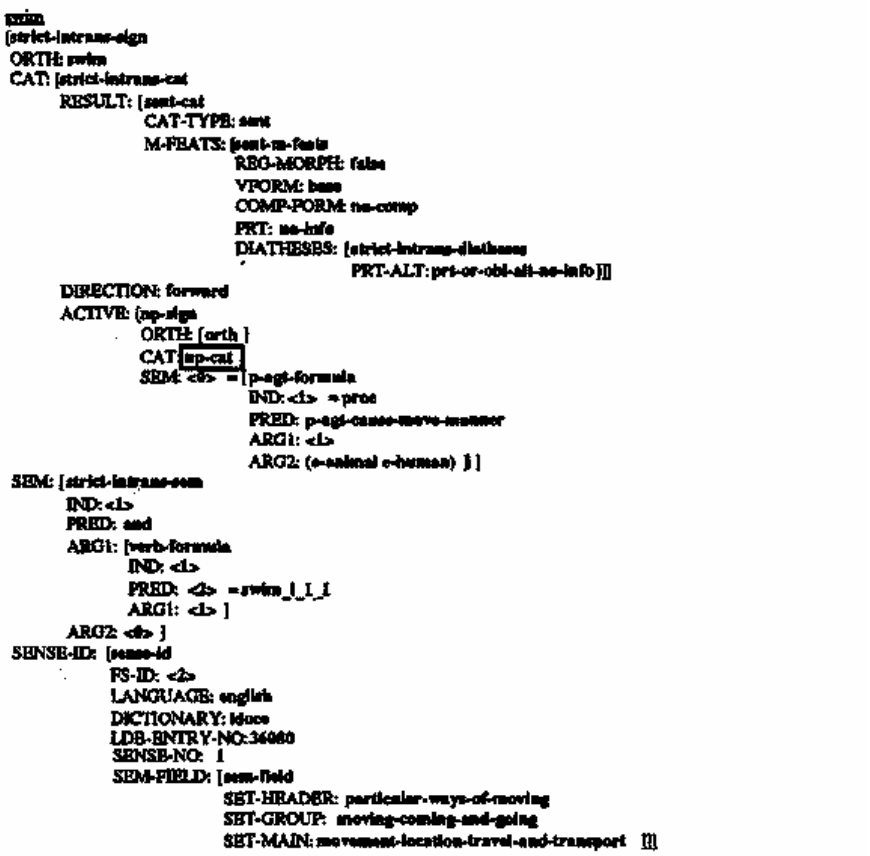

Figure 1: LKB entry for sense 1 of the verb swim in LDOCE.[4]

Consider, for example, the semantic classification of movement verbs adopted in the LKB. Following Talmy (1985), we assume that the semantic characterization of a movement situation involves reference to the following components: *causation, motion, path, moving object, reference location* (e.g. source, goal), and *manner* (of movement). These meaning components can be lexicalized independently of each other, or clustered into a variety of combinations. To represent these possibilities in the LKB, we used the meaning components **cause, move, manner, path, source, goal** to sort the thematic predicates used in verb representations. For example, the subject of *swim* in Figure 1 is associated with the participant role type **p-agt-cause-move-manner** indicating that self-causing, undirected movement for which manner is specified is involved. The same thematic specification can be used to characterize the semantics of *nuotare* and *nadar* which translate *swim* into Italian and Spanish respectively, as shown in Figure 2. Predicate sorting is also enforced for prepositional arguments. For example, locative prepositions expressing direction are distinguished as to whether they imply a completed path (e.g. *to, across)* or a path for which no end point is specified (e.g. *along, around).*

[4] According to the verb representation adopted in the LKB (Sanfilippo, 1992), verbs are treated as predicates of eventualities and thematic roles as relations between eventualities and individuals (Parsons, 1990). The semantic content of roles is computed in terms of entailments of verb meanings which determine the most (**P-AGT-..** .) and least **(P-PAT-.** . .) agentive event participants for each choice of predicate, and which are instrumental in providing a decompositional characterization of semantic verb classes. This approach reproduces the insights of Dowty's and Jackendoff's treatments of thematic information (Dowty, 1991; Jackendoff, 1990) within a neo-Davidsonian approach to verb semantics (Sanfilippo, 1990, 1992). A box around a type as in the case of np-cat in Figure 1 indicates that the feature structure associated with the type has been shrunk to ease graphical representation.

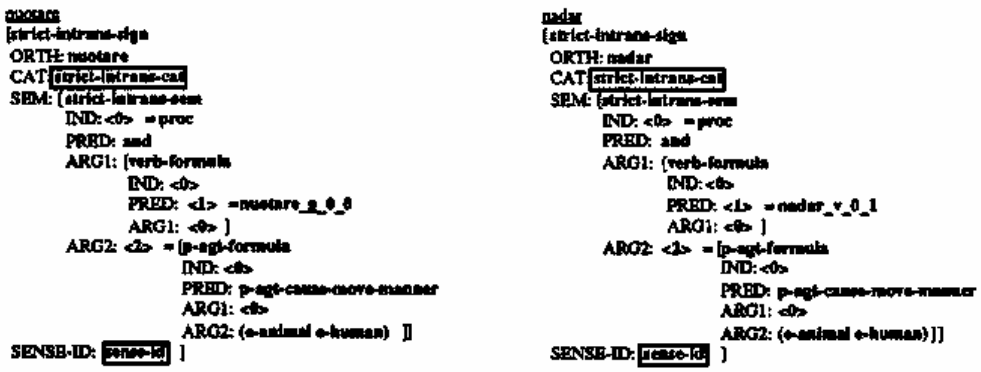

Figure 2: LKB entries for the Italian and Spanish translations of *swim*

In principle, the use of a common type system in the domain of semantic representation could be made to provide the kind of conceptual representation which is used in interlingual approaches to MT (Lytinen & Schank, 1982; Dorr, 1990). This is not the case, however, in the ACQUILEX LKB where semantic decomposition is only partially executed; for example, language-specific predicates (i.e. names of word senses such as **swim_1_1_1**) are still needed to differentiate word meanings. Consequently, our treatment makes it possible to exploit some of the advantages of an interlingual approach without a specific commitment to expressing all aspects of word meaning in terms of a language-independent semantic representation. While use of an interlingual semantic representation is appealing in that it should be the best solution to the problem of assessing translation equivalence between words, the task of providing such a specification reliably for large scale lexicons must remain a goal for future research.

3 Translation Equivalence

The *tlink* mechanism uses typed feature structures (FSs) to express translation equivalence between expressions in the source and target languages. More precisely, a *tlink* is defined as a FS which describes how lexical entries in the target and source languages can be made into translation equivalent pairs. In the simplest case, a *tlink* establishes a correspondence between two FSs which represent single (untransformed) lexical entries. In general, however, the *tlink* mechanism can relate sets of lexical items, and makes it possible to transform lexical entries into translationally equivalent pairs using the lexical or phrasal rules of the monolingual grammars. Such relationships are expressed concisely and can be used to state generalizations across classes of translation equivalences by exploiting the inheritance system on which the LKB is based via the types as well as the *psorts* which allow inheritance from lexical items and rules. The type **tlink** is defined as a bipartite structure encoding FS representations of lexical entries and possibly phrases (lexical and phrasal *signs)* for the source and target languages:

(2) tlink (top)
< sfs > = top-rule
< tfs > = top-rule.

The type **top-rule** in (2) — defined in (3c) as a subtype of **rule-or-sign-or-set-of-signs** — subsumes ordinary lexical and phrasal rules, or can be used to establish an equivalence between two (sets of) signs (sign-or-set-of-signs) of which one (or more) can be the result of a lexical/phrasal rule:[5]

(3) a rule-or-sign-or-set-of-signs (Top).
b sign-or-set-of-signs (rule-or-sign-or-set-of-signs).
c top-rule (rule-or-sign-or-set-of-signs)
< 0 > = sign-or-set-of-signs
< 1 > = rule-or-sign-or-set-of-signs.

[5] *Tlinks* are symmetrical and reversible; we use the terminology source *(***sfs***)*, target *(***tfs***)*, input (1) and output (**0**) solely for ease of exposition.

Minimally, a rule consists of one input FS and the output FS:

(4) rule (top-rule)
< 0 > = sign
< 1 > = sign.

With lexical rules, both input and output FSs describe lexical signs.

(5) lexical-rule (rule)
< 0 > - lex-sign
< 1 > - lex-sign.

For all *tlinks,* the FSs at the end of the output paths (< **sfs** : 0 > , < **tfs** : 0 >) will be translation equivalent when the input paths are instantiated by lexical entries. By defining types of *tlinks,* the concept of translation equivalence can be constrained and generalizations can be encoded. In a large variety of cases, translation equivalence can be straightforwardly expressed as a relation between untransformed lexical signs incorporating specified reentrant links between source and target FSs. For example, the *tlink* **simple-strict-intrans-tlink**, defined as subtype of **simple-tlink**, sets two lexical signs of type **strict-intrans-sign** to be directly equivalent — i.e. the input and output paths on both sides of the *tlink* share the same value; in addition, semantic indexes (i.e. the event and object variables relative to the top index and subject argument role) are made reentrant and thus constrained to be identical.

(6) a simple-tlink (tlink)
< sfs : 0 > = < sfs : 1 >
< tfs : 0 > = < tfs : 1 >
< sfs : 0 : sem : ind > = < sfs : 0 : sem : ind >.

b simple-strict-intrans-tlink (simple-tlink)
< sfs : 0 : sem : arg2 : arg2 > = < tsf : 1 : se» : arg2 : arg2 >.

Individual *tlinks* are created by instantiating the input paths by non-default inheritance (indicated by the symbol <=) from entries in the monolingual lexicons.

(7) swim_L_1_1/nadar_V_0_0
simple-strict-intrans-tlink
< sfs : 1 > <= swim_L_1_1 <>
< tfs : 1 > <= nadar_V_0_0 <>.

The expanded version of the *tlink* in (6) given in Figure 3 provides a concrete illustration of the equivalence discussed.

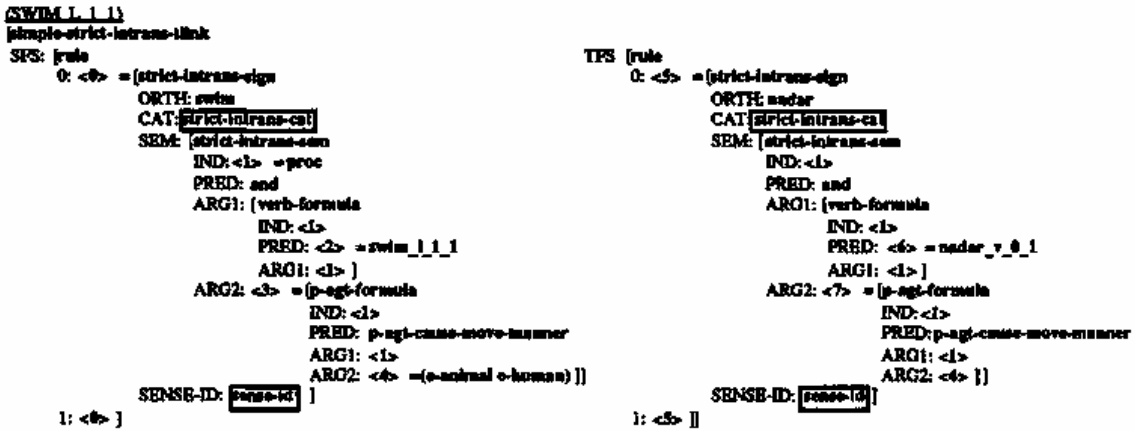

Figure 3: *tlink* for **swim_L_1_1** and **nadar_V_0_0**

At present, it will suffice to say that the reentrancies across the source and target templates in a *tlink* — e.g. <**1**> and <**4**> in Figure 3 — ensure that the constraints encoded by sorted variable (e.g. selectional restrictions) in the source and target signs are compatible. Their significance will become clearer in the next section where the transfer regime induced by the *tlink* mechanism is related to the translation task

as a whole. With more complex examples of translation equivalence, the binding of semantic indexes across the two sides of the *tlink* equation is instrumental in expressing translation mismatches. Consider, for example, the case of thematic divergence with reference to English/Italian translation pairs such as *like/piacere,* a very well-known example widely discussed in the literature on MT (Dorr, 1990; Tsujii, 1991; Beaven, 1992; Whitelock, 1992). In this case, the translation equivalence holds when the order of experiencer and stimulus argument variables is switched around in the verb's argument structure, as indicated by the reentrancies **<3>** and **<5>** in the expanded *tlink* declaration in Figure 4. This accounts for the distinct surface realizations of the experiencer and stimulus arguments in the two languages, cf. $Carlo_{subj}$ *likes* $Mary_{obj}$ translates into Italian $Mary_{subj}$ *piace a* $Carlo_{ind\ obj}$ (literally, 'Mary is pleasing to Carlo'). Moreover, the target output of the *tlink* includes an additional FS for the preposition which governs the experiencer argument (cf. *a* in the Italian example); the predicate name and semantic indexes of this preposition are instantiated by corresponding values for the indirect-object semantics in the argument structure of the verb, as indicated by the tag **<8>** in Figure 4.

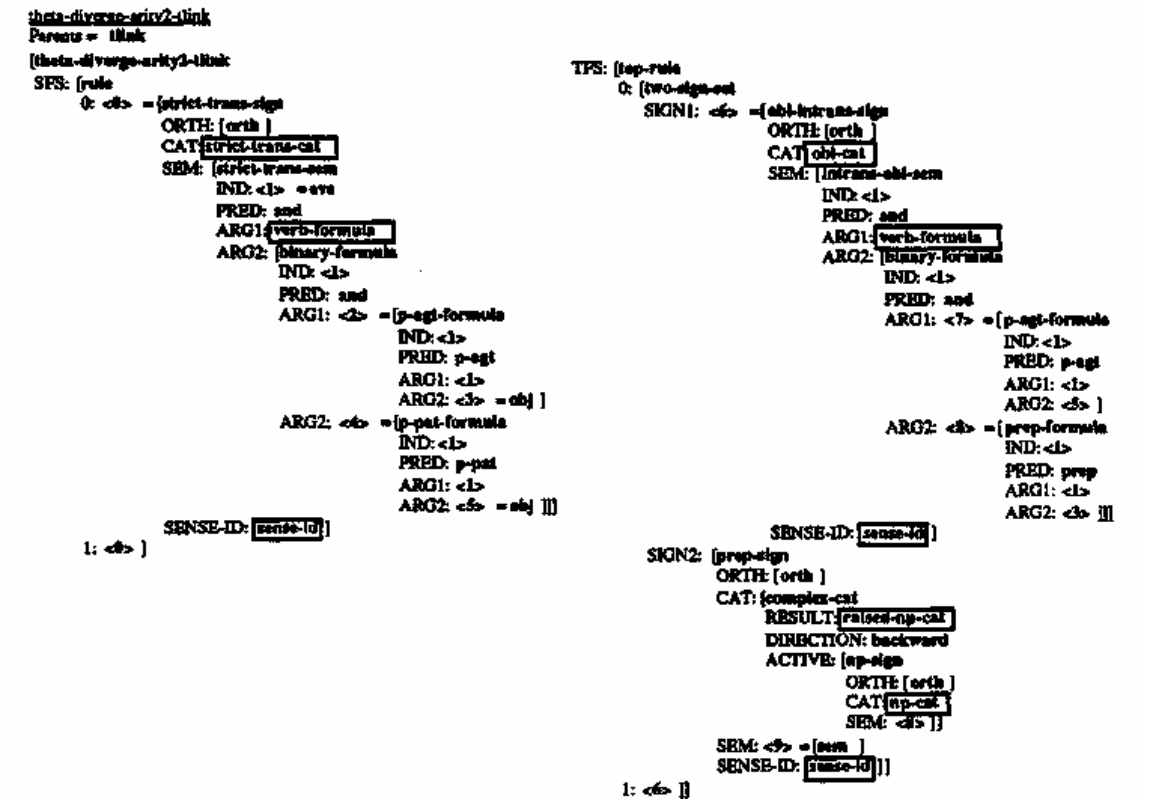

Figure 4: *tlink* for thematic divergence mismatches

The *tlink* mechanism can also integrate rule application both in the source and target sides of the transfer equation. Details about the use of rule application with *tlinks* are given in Copestake *et al.* (1992) where several instances including lexical and phrasal rules are discussed. For our purposes, it will suffice to discuss an example of complex transfer which makes use of lexical rule application. Consider again the translation mismatch discussed in the introduction with reference to movement verbs in the English/Spanish sentences:

(8) a John swam across the river

b John cruzó el río nadando

Given the semantic classification of movement verbs sketched in §2, the translation equivalence between *swim across NP*[6] and *cruzar NP nadando* could be expressed in terms of a *tlink* type which established

[6] This instance of *swim* can be generated from the sign in Figure 1 through a lexical rule which augments manner-of-

the following transfer pattern:

> An intransitive movement verb expressing manner of motion which subcategorizes for an directional argument implying a completed path together with the preposition which instantiates such a path (e.g. *swim across*) are translated into a verb of motion which subcategorizes for an expression of completed-path *(cruzar/alcanzar)* plus the translation equivalent of the source movement verb with its directional argument and path specification removed *(nadar):*

(9) $\left[\begin{array}{l}\textbf{Verb}_{Eng}\\ \textbf{Subj: } move\text{-}manner\text{-}path\\ \textbf{Obj: } goal_{across}\end{array}\right]\left[\begin{array}{l}\textbf{Prep}_{Eng}\\ goal_{across}\end{array}\right] \approx \left[\begin{array}{l}\textbf{Verb}_{Sp}\\ \textbf{Subj: } move\text{-}path\\ \textbf{Obj: } goal_{across}\end{array}\right]\left[\begin{array}{l}\textbf{Verb}_{Sp}\\ \textbf{Subj: } move\text{-}manner\end{array}\right]$

There are two major advantages in having a transfer pattern of this kind. First, the equivalence stated is linguistically motivated as it reflects crosslinguistic generalizations about lexicalization patterns with movement verbs (Talmy 1985; Jackendoff, 1990) which are reflected in grammatical processing (Levin & Rappaport, 1991). Second, because of its generality the equivalence can be easily extended to other movement verbs which exhibit the same complex behaviour — most manner-of-motion verbs, e.g. *amble, crawl, float, hobble, run, walk* — letting specific prepositions select a lexical instantiation (or set of possible instantiations) for the additional movement verb introduced in the target language. For example, *crawl across* would translate as *cruzar gateando,* while *subir gateando, descender gateando, alcanzar gateando* (literally 'ascend/descend/reach crawling') would be equivalent to *crawl up, crawl down, crawl to.* In addition, both *cruzar nadando* and *atraversar nadando* are translationally equivalent to swim *across* and it would therefore be appropriate to make either option available. Ultimately, the translation equivalence described would have extensions for each choice of preposition.

Needless to say, the informal characterization given in (9) fails to capture other important aspects of the transfer concerning correspondences between participant roles and the temporal interpretation of the two verbs in the target-language side of the equivalence. First of all, that the subject argument in the source language corresponds to the subject of the two verbs in the target language, e.g.

(10) X swim-across ≈ X cruzar & X nadar

Second, the directional argument (goal) of the verbs in the source and target languages coincide, e.g.

(11) X swim-across Y ≈ X cruzar Y & X nadar

Third, the events described by the verb introduced in the target language and the source verb are both amenable to telic interpretation as they imply a completed path;[7] by contrast, the verb which translates the source verb with its directional argument removed describes a process since it does not incorporate a path specification, e.g.

(12) X swim-across(dyn-eve) Y ≈ X cruzar(dyn-eve) Y & X nadar(proc)

Furthermore, the events described by the two verbs in the target language are co-extensive, i.e. if John crossed the river swimming (cf. *John cruzó el río nadando),* then it is necessarily the case that he was swimming while he did the crossing.

(13) X swim-across(dyn-eve) Y ≈ X cruzar(dyn-eve) Y & X nadar(proc) & while(proc, dyn-eve)

Such a temporal dependence can be induced by means of a lexical rule which turns the target manner-of-motion verb (e.g *nadar)* into a gerundive verb-phrase modifier where the events described by the the gerundive modifier and argument VP *(cruzar)* stand in the *while* relation. This rule can be introduced in the target language side of the *tlink* stating the equivalence for translation pairs such as *swim-across/nadar,* and its result co-bound with one of the output target-language signs. The expanded *tlink* below provides a detailed implementation of the equivalences described for the class of English/Spanish translation of movement verbs discussed.

---

movement verbs in English with a subcategorized prepositional phrase expressing completed path.

[7] More precisely, the events are 'dynamic' in that describe a movement situation but can be interpreted as either accomplishments/achievements (e.g. *John swam across the river in ten minutes)* or a process *(John swam across rivers all day).*

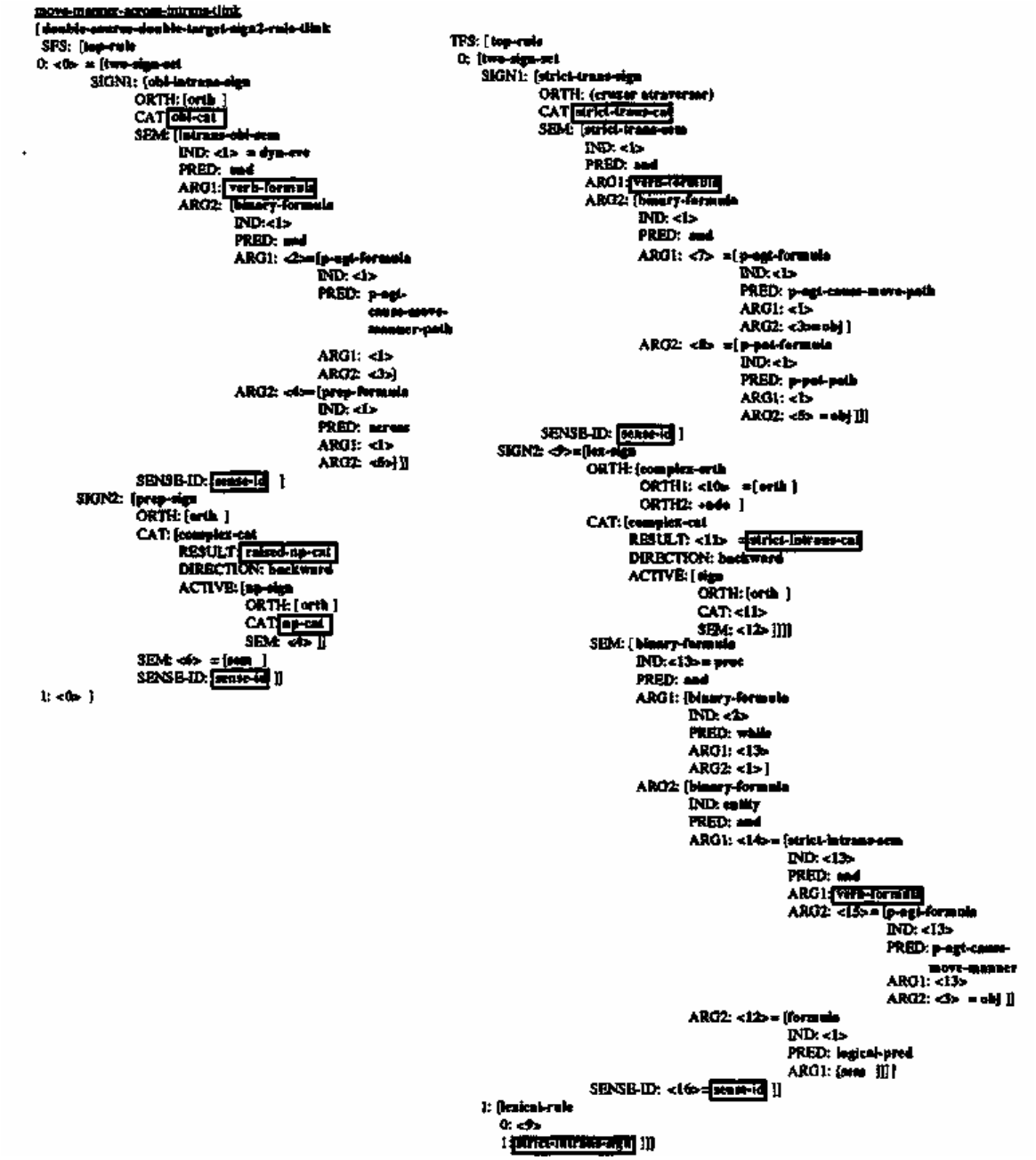

Figure 5: Complex *tlink* for head switching mismatches with movement verbs

4 Using *tlinks*

The *tlink* mechanism provides a conception of translation equivalence which is particularly suitable for transfer-based approaches to MT. Current systems which are attuned to this methodology (Kaplan et al., 1989; Estival et al., 1990), are wont to assume that the information structures which are passed on to the generation component are structured according to the source-language parse appropriately modified

through transfer rules which establish translation equivalences. Concerning the treatment of complex transfer such as the 'head switching' mismatches discussed above, this practice requires the creation of rules which make it possible to rearrange grammatical dependencies involving constituents other than those included in the same lexical equivalence. In the translation pair *swim across the river ≈ cruzar el río nadando,* for example, lexical transfer involves three distinct tlinks *(swim across ≈ cruzar nadando, the ≈ el, river ≈ río*) while a single structural equivalence arches over all phrasal constituents (V Prep NP ≈ V NP V). Because of its strong lexicalist orientation, the *tlink* mechanism is not equipped to express structural equivalences which arch over lexical entries involved in distinct *tlinks.* For example, we could design a rule which made the two target output signs in the *tlink* for 'head-switching' mismatches in Figure 5 amenable to combination. This would make it possible to create a single information structure corresponding to the verbal complex *cruzar nadando;* however, if we were to use such an expression as input to generation, English *swim across the river* would translate as *cruzar nadando el río* which is not what we want.

Needless to say, this generation problem could be solved by allowing syntactic information to bear on the functionality of *tlinks* so that information about the parse tree in the source language would also become input to translation equivalence. This practice, however, would yield equivalences between lexically governed syntactic rules which are unmotivated from the perspective of monolingual grammars. It would therefore be desirable to maintain the current lexicalist orientation and find a way to relate information about the target language to generation which does not rely on transfer of the parse tree in the source language. Such an alternative has been recently explored in detail by Whitelock (1992) and Beaven (1992) within a novel approach to machine translation which has come to be known as *Shake and Bake* (S&B).

As in the *tlink* mechanism described in this paper, transfer in S&B MT is only meant to provide a specification of lexical equivalence. During translation, the lexical entries resulting from the analysis of the source language string provide input to transfer templates similar to our *tlinks* to yield a set of translationally equivalent lexical entries in the target language. These are then combined freely using the target language grammar. The regime of variable sharing across the source and target languages enforced during lexical transfer (as in our *tlinks)* provides the necessary constraints to ensure the invariance of semantic dependencies in translation. Consider, for example, the translation of the Italian/English pair of nominals in (14) using the monolingual rules in (15)-(16).

(14) a libri francesi e giornali
b French books and journals

(15) *Italian Grammar*

1. $\begin{bmatrix} w1\ w2\ w3 \\ N \\ [z]\,[A\ \&\ B\ \&\ group(z,x,y)] \end{bmatrix} \rightarrow \begin{bmatrix} w1 \\ N \\ [x]\,A \end{bmatrix} \begin{bmatrix} w2 \\ conj \end{bmatrix} \begin{bmatrix} w3 \\ N \\ [y]\,B \end{bmatrix}$

2. $\begin{bmatrix} w1\ w2 \\ N \\ [x]\,[A\ \&\ P] \end{bmatrix} \rightarrow \begin{bmatrix} w1 \\ N \\ [x]\,A \end{bmatrix} \begin{bmatrix} w2 \\ ADJ \\ [x]\,P \end{bmatrix}$

(16) *English Grammar*

1. $\begin{bmatrix} w1\ w2\ w3 \\ N \\ [z]\,[A\ \&\ B\ \&\ group(z,x,y)] \end{bmatrix} \rightarrow \begin{bmatrix} w1 \\ N \\ [x]\,A \end{bmatrix} \begin{bmatrix} w2 \\ conj \end{bmatrix} \begin{bmatrix} w3 \\ N \\ [y]\,B \end{bmatrix}$

2. $\begin{bmatrix} w1\ w2 \\ N \\ [x]\,[A\ \&\ P] \end{bmatrix} \rightarrow \begin{bmatrix} w1 \\ ADJ \\ [x]\,P \end{bmatrix} \begin{bmatrix} w2 \\ N \\ [x]\,A \end{bmatrix}$

Because of differences in word order, the scope of the adjective is unambiguous in Italian but not in English. If translating from Italian, we would thus like to obtain a single reading — i.e. one where the adjective modifies the adjacent noun only (e.g. *libri/books).* S&B MT can be made to provide the desired

result without using the source parse tree as input to generation. First, the source string is parsed using the Italian grammar giving as result the phrasal nominal in (17).

(17) $\left[\begin{array}{l} \textbf{libri francesi e giornali} \\ N \\ [z]\,\{l(x)\ \&\ f(x)\ \&\ g(y)\ \&\ group(z,x,y)\} \end{array}\right]$

Then, the lexical entries used in the successful parse are allowed to instantiate the source language side of transfer templates *(tlinks);* before this instantiation takes place, each variable is replaced by a unique constant to preserve the semantic dependencies established in the parse:

(18)
$$\left[\begin{array}{l} \textbf{libri} \\ N \\ [\mathrm{SK}_x]\,l(\mathrm{SK}_x) \end{array}\right] \quad \left[\begin{array}{l} \textbf{francesi} \\ ADJ \\ [\mathrm{SK}_x]\,f(\mathrm{SK}_x) \end{array}\right] \quad \left[\begin{array}{l} \textbf{e} \\ Conj \end{array}\right] \quad \left[\begin{array}{l} \textbf{giornali} \\ N \\ [\mathrm{SK}_y]\,g(\mathrm{SK}_y) \end{array}\right]$$
$$\left\{\left[\begin{array}{l} \textbf{books} \\ N \\ [\mathrm{SK}_x]\,b(\mathrm{SK}_x) \end{array}\right] \quad \left[\begin{array}{l} \textbf{French} \\ ADJ \\ [\mathrm{SK}_x]\,f(\mathrm{SK}_x) \end{array}\right] \quad \left[\begin{array}{l} \textbf{and} \\ Conj \end{array}\right] \quad \left[\begin{array}{l} \textbf{journals} \\ N \\ [\mathrm{SK}_y]\,j(\mathrm{SK}_y) \end{array}\right]\right\}$$

The lexical entries in the target language set are then combined in whichever sequential and hierarchical order is allowed by the English grammar rules in (16). A priori, there are many possible ways in which the four lexical entries can be combined; in practice, however, most will be ruled out through the constraints imposed by shared semantic indexes. For example, the adjective must combine with (and thus be adjacent to) the noun *books* which is the only expression to have a compatible semantic index (i.e. SKX). The semantic dependencies of the parse tree in the source language is thus preserved even though no structural transfer has been enforced.

A strongly lexicalist approach to translation equivalence such as that enforced by our *tlink* mechanism is fully compatible with an approach to MT similar to S&B. This is simply because the generation task is essentially reduced to parsing all possible sequences of lexical items in the target-language set, and there is no need for phrasal transfer to drive generation.

5 Conclusions

What has come to be called the lexical knowledge 'bottleneck' is now a major focus of research activity in NLP and MT as researchers have come to realise that prototype systems can only be transformed into useful applications once we have a methodology for the construction of substantial multilingual lexica in a resource efficient fashion. The methodology we propose draws on both the empiricist and rationalist traditions. On the one hand, the efficient creation of large scale lexicons requires application of (semi)automatic and statistical techniques for the acquisition of lexical knowledge from corpora and machine-readable dictionaries; this is in keeping with an empiricist orientation. On the other hand, the expressiveness of lexical descriptions is certainly determined by our understanding of theoretical issues in lexical semantics. For example, the treatment of translation equivalence can achieve both linguistic appropriateness and engineering efficiency if informed by a specification of crosslinguistic variation in lexicalization patterns. In this respect, the field of MT can benefit from theories of word meaning and language structure. In this paper, we have described a concrete example of how these two perspectives can be integrated. The eventual success and applicability of the specific approach we have proposed here rests on the development of computationally efficient versions of the Shake and Bake generation algorithm; and, of course, further linguistic refinement of *tlink* specifications. Nevertheless, the general methodology proposed here will, we believe, provide an appropriate framework in which to undertake such work.

References

Ageno, A., I. Castellón, M.A. Marti, F. Ribas, G. Rigau, H. Rodriguez, M. Taulé and F. Verdejo. From LDB to LKB. ESPRIT BRA-3030 ACQUILEX - WP No. 039, 1992.

Beaven, J. *Lexicalist Unification-Based Machine Translation,* PhD thesis, University of Edinburgh, 1992.
Beaven, J. & Whitelock, P. Machine Translation Using Isomorphic UCGs. In *Proceedings of COLING-88,* Budapest, 1988.
Briscoe, T. Lexical Issues in Natural Language Processing. In Klein, E. &; F. Veltman (eds.). *Natural Language and Speech,* Springer-Verlag, pp. 39-68, 1991.
Briscoe, T., A. Copestake and V. de Paiva (eds.) *Default Inheritance within Unification-Based Approaches to the Lexicon.* Cambridge University Press, forthcoming.
Calzolari, N. Acquiring and Representing Semantic Information in a Lexical Knowledge Base. In *Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation,* 1991.
Copestake, A. The ACQUILEX LKB: Representation Issues in Semi-Automatic Acquisition of Large Lexicons. In *Proceedings of the 3rd Conference on Applied Natural Language Processing,* 1992.
Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni, E. Marinai. Multilingual Lexical Representation. ESPRIT BRA-3030 ACQUILEX - WP No. 043.
Dorr, B. Solving Thematic Divergences in Machine Translation. In *Proceedings of the 28th Conference of the Association for Computational Linguistics,* 1990.
Dowty, D. Grammatical Relations and Montague Grammar. In Jacobson, P. and Pullum, G. K. (eds.) *The Nature of Syntactic Representation,* pp. 79-130. Dordrecht: D. Reidel. 1982.
Dowty, D. Thematic Proto-Roles and Argument Selection. Language 67, pp. 547-619, 1991.
Estivai, D., A. Ballim, G. Russell and S. Warwick. A Syntax and Semantics for Feature Structure Transfer. In *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in MT of NLs,* Austin, Texas, 1990.
Jackendoff, R. *Semantic Structures.* MIT Press, Cambridge, Mass, 1990.
Kaplan, R., K. Netter, J. Wedekind and A. Zaenen. Translation by Structural Correspondences. In *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics,* Manchester, 1989.
Levin, B. & Rappaport, M. The Lexical Semantics of Verbs of Motion: The Perspective from Unaccusativity. To appear in Roca, I. (ed) *Thematic Structure: Its Role in Grammar,* Foris, 1991.
Lytinen, S. & Schank, R. *Representation and Translation.* Technical Report 234, Department of Computer Science, Yale University, New Haven, CT, 1982.
McArthur, T. (1981) *Longman Lexicon of Contemporary English.* Longman, London.
Parsons, T. *Events in the Semantics of English: a Study in Subatomic Semantics.* MIT Press, 1990.
Procter, P. (1978) *Longman Dictionary of Contemporary English.* Longman, London.
Sanfilippo, A. *Grammatical Relations, Thematic Roles and Verb Semantics.* PhD thesis, Centre for Cognitive Science, University of Edinburgh, Scotland, 1990.
Sanfilippo, A. LKB Encoding of Lexical Knowledge. To appear in Briscoe, T., A. Copestake and V. de Paiva (eds.) *Default Inheritance within Unification-Based Approaches to the Lexicon.* Cambridge University Press, 1992.
Sanfilippo, A & V. Poznański. The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing,* Trento, 1992.
Talmy, L. Lexicalization Patterns: Semantic Structure in Lexical Form. In Shopen, T. (ed) *Language Typology and Syntactic Description 3. Grammatical Categories and the Lexicon,* CUP, 1985.
Tsujii, J. & Fujita, K. Lexical Transfer on Bilingual Signs: Towards Interaction during Transfer. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics,* Berlin, 1991.
Vossen, P. An Empirical Approach to Automatically Construct a Knowledge Base from Dictionaries. In *Proceedings of the 5th EURALEX,* Tampere, Finland, 1992.
Whitelock, P. The Organization of a Bilingual Lexicon. DAI Working Paper, Dept. of Artificial Intelligence, University of Edinburgh, 1988.
Whitelock, P. Shake-and-Bake Translation. To appear in *Proceedings of COLING-92.*
Zajac, R. A Transfer Model Using a Typed Feature Structure Rewriting System with Inheritance. In *Proceedings of the 27th Conference of the Association for Computational Linguistics,* 1989.