ACQUILEX II — Esprit Project 7315

Working paper number: 3

Date: January 1993

Title: Multilingual Lexical Representation

Authors: Ann Copestake and Antonio Sanfilippo
Computer Laboratory, University of Cambridge

In: 'Building Lexicons for Machine Translation', edited by Bonnie Dorr, Proceedings of the AAAI Spring Symposium, Stanford, California, 1993

Multilingual Lexical Representation

Ann Copestake and Antonio Sanfilippo

Abstract

The approach to multilingual lexical representation developed as part of the ACQUILEX Lexical Knowledge Base (LKB) is discussed with specific reference to complex translation equivalence. The treatment described provides a lexicalist account of translation mismatches in terms of translation links which capture cross-linguistic generalizations across sets of semantically related lexical items, and can be readily integrated with several transfer-based MT systems.

1 Introduction

The ACQUILEX LKB system was designed to allow the representation of syntactic and semantic information which has been (semi-)automatically extracted from machine readable dictionaries (MRDs). Large scale monolingual lexicon fragments have been constructed semi-automatically for four languages (English, Spanish, Dutch and Italian); descriptions of the monolingual lexicons and the lexical representation language (LRL) are given in, for example, Copestake (1992), Sanfilippo and Poznanski (1992) and papers in Briscoe *et al.* (in press). Here we describe our use of the LRL to represent multilingual information in the form of links between monolingual lexical entries, and discuss how this representation is usable by MT systems. We regard MT as only one possible application; ultimately the multilingual LKB should be able to support the needs of linguists, lexicographers and others who need access to substantial quantities of cross-linguistic information.

Recently, there has been considerable interest in using unification based formalisms to model transfer approaches to MT, aiming for declarativeness and bidirectionality, but allowing sufficient expressiveness to deal with complex classes of translation equivalence (e.g. Kaplan *et al.*, 1989; Zajac, 1989; Estival *et al.*, 1990; Alshawi *et al.*, 1991). We have attempted to maintain these advantages, but to abstract away from the aspects of these systems which are specific to particular MT techniques. This makes it possible to maximize the functionality of the system with respect to the expression of linguistic and lexicographic generalisations and facilitates the construction of a multilingual lexicon which would support a variety of approaches to MT, although it is naturally most appropriate to the more lexicalist frameworks.

Like parsing and generation, MT can be regarded very generally as a process of constraint solving. The target language (TL) sentence generated is constrained by the monolingual grammar and lexicon and also by constraints derived from the parse of the source language (SL) sentence. Different approaches to MT can be characterised according to what sort of cross-linguistic constraints are imposed, in addition to the monolingual constraints. The archetypal use of an interlingua corresponds to the situation where the structure produced by the parse of the SL sentence is identical to that which would be produced by the parse of the TL sentence. Transfer based approaches, in contrast, do not assume identity between any part of the SL structure and TL structure, but they differ with respect to the sort of correspondences which can be described. Just as grammar rules expressed in a unification-based formalism can be regarded non-procedurally as constraints on well-formed monolingual structures, transfer rules or translation links can be regarded as constraints on the mappings between source and target structures which are translation equivalent. The particular approach to translation which we are exploring is heavily lexicalist in orientation, although it is also possible to incorporate non-lexical constraints.

We define lexical translation equivalence in terms of cross-linguistic links, *tlinks*, stated in terms of lexical entries in the monolingual lexicons. Because tlinks are defined in terms of inheritance from lexical entries and rules, translation equivalence can be stated through direct reference to properties of word syntax and semantics as they appear in the monolingual lexical descriptions. The sharing of information structures between tlinks and the lexicons provides an efficient and linguistically motivated way of characterizing classes of lexical correspondence across languages, promotes compactness in the description of translation equivalence and ensures that the multilingual and monolingual components are compatible.

The LRL is designed to be a general representation language, capable of encoding a variety of linguistic approaches, following the same sort of philosophy as PATR-II (Shieber, 1986). This flexibility also applies to the tlink mechanism. In the LRL the encoding of a linguistic theory is expressed in the type system. In the case of the type system adopted for work on ACQUILEX, an important element of our approach is the adoption of a common system of types to encode syntactic and semantic properties of lexical items in the four languages investigated within the project. This ensures compatibility of representation without ruling out language specific parametrisation where necessary. Such a practice is particularly effective in the treatment of translation mismatches — see below and Sanfilippo *et al.* (1992). One other main feature of our approach to representation is our emphasis on representing detailed information about lexical semantics. Again, this is justified by monolingual considerations, but we exploit lexical semantic information in the multilingual aspects of representation.

2 Monolingual representation

Our approach to monolingual lexical representation involves combining syntactic, formal semantic and lexical semantic information in a single lexical entry. Lexical entries are represented as *typed feature structures* (FSs) where syntactic and semantic information is expressed in terms of attribute-value pairs. Consider, for example, the lexical entry for the noun sense for *(drinking) chocolate*,

$$\left[\begin{array}{l} \textbf{lex-noun-sign} \\ \text{ORTH} = \textbf{chocolate} \\ \text{CAT} = \boxed{\textbf{noun-cat}} \\ \text{SEM} = \left[\begin{array}{l} \textbf{unary-formula-entity-arg1} \\ \text{PRED} = \textbf{chocolate\_L\_1\_4} \\ \text{ARG1} = \textbf{entity} \end{array}\right] \\ \text{SENSE-ID} = \boxed{\textbf{sense-id}} \\ \text{QUALIA} = \left[\begin{array}{l} \textbf{c\_art\_subst} \\ \text{PURPOSE} = \left[\begin{array}{l} \textbf{formula} \\ \text{PRED} = \textbf{drink\_L\_1\_1} \end{array}\right] \\ \text{PHYSICAL\_STATE} = \textbf{liquid\_a} \\ \text{CONSTITUENCY} = \boxed{\textbf{constituency}} \\ \text{FORM} = \left[\begin{array}{l} \textbf{physform} \\ \text{SHAPE} = \textbf{non-individuated} \end{array}\right] \end{array}\right] \end{array}\right]$$

Figure 1: Lexical entry for (drinking) *chocolate*

shown below, which was derived automatically from the *Longman Dictionary of Contemporary English* (LDOCE; Procter, 1978). Here **lex-noun-sign** is a type which specifies syntactic and semantic properties of nouns and `drink_L_2_1` is the name for the lexical entry from which a value for the QUALIA attribute is inherited by default (`<`). The remainder of the entry specifies the MRD source.

```
chocolate_L_1_4
lex-noun-sign
<SENSE-ID : DICTIONARY> = "LDOCE"
<SENSE-ID : HOMONYM-NO> = "1"
<SENSE-ID : SENSE-NO> = "4"
<SENSE-ID : LDB-ENTRY-NO> = "5902"
<QUALIA> < drink_L_2_1 <QUALIA> .
```

When expanded, the lexical description above will yield a FS containing syntactic and semantic information, as partially shown in AVM notation in Figure 1. (Throughout this paper we will use bold font for types and capitals for features. In the AVM notation, reentrancy is indicated by integer tags, angle brackets denote list structures and the boxes indicate parts of the FS which are not shown completely.) Our use of type inheritance and default inheritance allows a compact lexical entry to be expanded into a detailed FS. This makes it possible to avoid redefining the same information structures, thus reducing a great deal of redundancy in the specification of word forms. The syntactic and formal semantic portions of the lexical sign are relatively conventional, however we also encode detailed lexical semantic information, the QUALIA structure, in a way which is loosely based on Pustejovsky's work (e.g. Pustejovsky, 1991). The representation language used for lexical description is also employed for grammar and lexical rules, which are expressed as typed feature structures describing relationships between two or more signs (FS descriptions of single words or phrases). We will use a simplified version of the type system for the examples in this paper for ease of exposition.

3 Representation of translation equivalence

Tlinks can be viewed as information structures describing ways in which input lexical entries from the source and target languages are mapped into output translation-equivalent pairs. All tlinks are defined as FSs of type **tlink** which relates the source FS (SFS) and the target FS (TFS) which are both of type **rule**:

```
tlink (top)
< SFS > = rule
< TFS > = rule.
```

Minimally, a rule establishes a correspondence between an input sign (1) and an output sign (0):

```
rule (top)
< 0 > = sign
< 1 > = sign.
```

The rule-inputs of tlinks are meant to be instantiated by FS representations of word senses in the source and target languages; rule-outputs provide the translation equivalence. Thus, when the FSs at the end of the paths `< SFS : 1 >` and `< TFS : 1 >` are instantiated by lexical entries, the FSs at the end of the paths `< SFS : 0 >` and `< TFS : 0 >` are defined to be translation equivalent.[1] This level of indirection is crucial in expressing translation mismatches. Tlinks can be regarded as generating new FSs; given a FS in one language, and an appropriate tlink, unification with the FS at the end of the appropriate path, the *SL output*, (e.g. `< SFS : 0 >`) in the tlink, will result in the FS at the end of the other output path (the *TL output*) being returned (e.g. `< TFS : 0 >`).

The concept of translation equivalence is constrained by defining an inheritance network of tlink types encoding generalisations relative to classes of crosslinguistic equivalences. The commonest and simplest cases of translation equivalence can be represented as **simple-tlink**s.

[1] Tlinks are both symmetrical and reversible; we use the terminology source, target, input and output solely for ease of exposition.

```
simple-tlink (tlink)
< SFS : 0 > = < SFS : 1 >
< TFS : 0 > = < TFS : 1 >
< SFS : 0 : SEM : ARG1 > =
   < TFS : 0 : SEM : ARG1 >.
```

A **simple-tlink** is applicable in the case where two lexical entries which denote single place predicates (nouns etc) are straightforwardly translation equivalent, without any transformation being necessary. The semantic variables relative to the semantic entity described by two output structures are specified to be identical. This variable equivalence is the basis for the use of tlinks within unification based approaches to MT and is crucial in expressing translation mismatches such as thematic divergence and head switching (Sanfilippo *et al.*, 1992). The particular paths equated will depend on the way that the type system is used to encode semantic structure — in the examples which follow a simplified encoding has been used for ease of exposition.

For example, assuming that the LDOCE sense *chocolate* 1 4, is translation equivalent to the Van Dale *chocolade* 0 2, we would have the tlink:

```
simple-tlink
< SFS : 1 > <= chocolate_L_1_4 <>
< TFS : 1 > <= chocolade_V_0_2 <>.
```

where `<=`, indicates non-default inheritance from a named FS (the particular lexical entries). This tlink will yield the English entry as output when the Dutch entry is supplied as input and vice versa. The 'syntactically sugared' notation for this tlink, which will be used in subsequent examples, is:

```
chocolate_L_1_4 / chocolade_V_0_2
simple-tlink.
```

For intransitive verbs we will make use of a slightly different tlink:

```
iverb-tlink (tlink)
< SFS : 0 > = < SFS : 1 >
< TFS : 0 > = < TFS : 1 >
< SFS : 1 > = iverb-sign
< TFS : 1 > = iverb-sign
< SFS : 0 : SEM : ARG2 > =
  < TFS : 0 : SEM : ARG2 >.
```

For example the tlink for 'English:*fly* $\approx$ Italian:*volare*' would be:

```
fly_L_1_1 / volare_G_0_1
iverb-tlink.
```

$$
\left[\begin{array}{l}
\textbf{iverb-tlink} \\
\text{SFS} = \left[\begin{array}{l}
\textbf{rule} \\
0 = \boxed{0}\left[\begin{array}{l}
\textbf{iverb-sign} \\
\text{ORTH} = \langle \textbf{flies} \rangle \\
\text{CAT} = \boxed{\textbf{ivcat}} \\
\text{SEM} = \left[\begin{array}{l}
\textbf{ivsem} \\
\text{PRED} = \textbf{and} \\
\text{ARG1} = \left[\begin{array}{l}
\textbf{vformula} \\
\text{PRED} = \textbf{fly\_L\_1\_1} \\
\text{ARG1} = \boxed{2}\,\textbf{process}
\end{array}\right] \\
\text{ARG2} = \boxed{1}\left[\begin{array}{l}
\textbf{thformula} \\
\text{PRED} = \textbf{move-manner} \\
\text{ARG1} = \boxed{2} \\
\text{ARG2} = \textbf{obj}
\end{array}\right]
\end{array}\right]
\end{array}\right] \\
1 = \boxed{0}
\end{array}\right] \\
\text{TFS} = \left[\begin{array}{l}
\textbf{rule} \\
0 = \boxed{3}\left[\begin{array}{l}
\textbf{iverb-sign} \\
\text{ORTH} = \langle \textbf{vola} \rangle \\
\text{CAT} = \boxed{\textbf{ivcat}} \\
\text{SEM} = \left[\begin{array}{l}
\textbf{ivsem} \\
\text{PRED} = \textbf{and} \\
\text{ARG1} = \left[\begin{array}{l}
\textbf{vformula} \\
\text{PRED} = \textbf{volare\_G\_0\_1} \\
\text{ARG1} = \boxed{2}
\end{array}\right] \\
\text{ARG2} = \boxed{1}
\end{array}\right]
\end{array}\right] \\
1 = \boxed{3}
\end{array}\right]
\end{array}\right]
$$

Figure 2: Expanded tlink for 'English:*fly* ≈ Italian:*volare*'

The expanded version of this tlink is shown in Figure 2 where the type **move-manner** specifies the thematic functionality of the subject participant as involving movement with manner of motion expressed — see Sanfilippo (in press) for details about the representation of verb semantics adopted.

Tlinks can be viewed as constraining the relationship between structures in the source and target languages. If the SL output of some tlink unifies with some part of the structure that results from parsing a sentence, then the structure given to the TL sentence is constrained to include the TL output of that instantiated tlink. Consider, for example, the structure in Figure 3 which is a simplified version of the analysis for the sentence *Tweety flies* (ignoring morphology). Here we have assumed an HPSG-like treatment: note that the sign is equivalent to an entire parse tree, in that it contains the signs that have been combined to form the sentence sign itself, and that the arguments of the lexical signs have been coindexed as a result of the parse. The SL output of the tlink shown in Figure 2 can be unified with the instantiated lexical sign for the verb (i.e. the FS value for the path `< DTRS : HEAD-DTR >` in Figure 3) and therefore the TL structure is constrained to contain a structure which is subsumed by the TL output of the tlink. The same applies to the tlink relating the SL and TL subject noun phrases which in this case expresses simple identity. Because the variables corresponding to the arguments of the lexical sign are coindexed in the SL structure, and the tlinks contain statements of argument

$$
\begin{bmatrix}
\textbf{s-sign} \\
\text{ORTH} = \langle \boxed{1}\ \textbf{Tweety},\ \boxed{2}\ \textbf{flies} \rangle \\
\text{CAT} = \boxed{\textbf{scat}} \\
\text{SEM} = \begin{bmatrix}
\text{PRED} = \textbf{and} \\
\text{ARG1} = \boxed{3} \begin{bmatrix}
\textbf{ivsem} \\
\text{PRED} = \textbf{and} \\
\text{ARG1} = \begin{bmatrix} \textbf{vformula} \\ \text{PRED} = \textbf{fly\_L\_1\_1} \\ \text{ARG1} = \boxed{4}\,\textbf{process} \end{bmatrix} \\
\text{ARG2} = \begin{bmatrix} \textbf{thformula} \\ \text{PRED} = \textbf{move-manner} \\ \text{ARG1} = \boxed{4} \\ \text{ARG2} = \boxed{5}\ \textbf{obj} \end{bmatrix}
\end{bmatrix} \\
\text{ARG2} = \boxed{6} \begin{bmatrix} \textbf{npsem} \\ \text{PRED} = \textbf{tweety} \\ \text{ARG1} = \boxed{5} \end{bmatrix}
\end{bmatrix} \\
\text{DTRS} = \begin{bmatrix}
\textbf{head-comp-struc} \\
\text{HEAD-DTR} = \begin{bmatrix}
\textbf{iverb-sign} \\
\text{ORTH} = \langle \boxed{2} \rangle \\
\text{CAT} = \begin{bmatrix} \textbf{ivcat} \\ \text{HEAD} = \textbf{verb} \\ \text{SUBCAT} = \langle \boxed{7} \rangle \end{bmatrix} \\
\text{SEM} = \boxed{3}
\end{bmatrix} \\
\text{COMP-DTRS} = \left\langle \boxed{7} \begin{bmatrix} \textbf{np-sign} \\ \text{ORTH} = \langle \boxed{1} \rangle \\ \text{CAT} = \boxed{\textbf{npcat}} \\ \text{SEM} = \boxed{6} \end{bmatrix} \right\rangle
\end{bmatrix}
\end{bmatrix}
$$

Figure 3: Structure resulting after parse of source language sentence

equivalence, the arguments of the lexical signs in the TL structure will also be constrained to be identical. Given the constraints of the monolingual grammar, plus the additional assumption that no additional predicates may be inserted, this is sufficient information to constrain the TL structure for the complete sentence to that shown in Figure 4.

$$\left[\begin{array}{l} \textbf{s-sign} \\ \text{ORTH} = \langle \textbf{Tweety, vola} \rangle \\ \text{CAT} = \boxed{\textbf{scat}} \\ \text{SEM} = \left[\begin{array}{l} \text{PRED} = \textbf{and} \\ \text{ARG1} = \left[\begin{array}{l} \textbf{ivsem} \\ \text{PRED} = \textbf{and} \\ \text{ARG1} = \left[\begin{array}{l}\textbf{vformula}\\ \text{PRED} = \textbf{volare\_G\_0\_1} \\ \text{ARG1} = \boxed{4}\,\textbf{process}\end{array}\right] \\ \text{ARG2} = \boxed{1}\left[\begin{array}{l}\textbf{thformula}\\ \text{PRED}=\textbf{move-manner}\\ \text{ARG1}=\boxed{4}\\ \text{ARG2}=\boxed{5}\ \textbf{obj}\end{array}\right]\end{array}\right] \\ \text{ARG2} = \left[\begin{array}{l}\textbf{npsem}\\ \text{PRED}=\textbf{tweety}\\ \text{ARG1}=\boxed{5}\end{array}\right]\end{array}\right] \\ \text{DTRS} = \boxed{\textbf{head-comp-struc}} \end{array}\right]$$

Figure 4: Italian translation for *Tweety flies*

We have implemented a general constraint-based system which basically works by attempting to match tlinks against the SL structure resulting from a parse and then generating from the "bag" of feature structures in the TL which result. This has strong similarities with the Shake-and-Bake approach to translation (Beaven, 1992; Whitelock, 1992). However, non-lexical constraints may also be incorporated. For example, we might wish to ensure that the speech act type represented by the SL and TL sentences (e.g. *wh*-question, *yes/no*-question or imperative) match. In our MT system, this can be done by stating a constraint on the **s-sign**s, without any reference to lexical entries. The system is intended for testing tlinks, rather than as a practical approach to MT, and it is therefore designed for generality rather than efficiency. Further details of this system, including descriptions of how it handles multiple possibilities and phrasal equivalences, are given in Copestake (1993). We will consider the relationship to more practical approaches to MT in the conclusion, and concentrate on the lexical aspects here.

4 More complex translation links

Some restrictions on translation can be expressed by making the target or source FSs of tlinks more specific. For example, we can define a type **human-tlink** and state as a constraint that the values for the SEX feature must be the same in the translation equivalent feature structures:

```
human-tlink (simple-tlink)
< SFS : 0 : QUALIA : SEX > =
   < TFS : 0 : QUALIA : SEX >.
```

A tlink of this type would be suitable for establishing equivalences such as that between the word *teacher* in English and its translation in Spanish as either *maestro* or *maestra*. The restriction that *maestro* denotes a male teacher and *maestra* a female one — i.e. the values for `< QUALIA : SEX >` in the FS descriptions for the two words are **male** and **female** respectively — follows from the use of tlinks of type **human-tlink** to relate *teacher* with *maestro* and *maestra*:

```
teacher / maestro :
human-tlink.

teacher / maestra :
human-tlink.
```

The path equation induced by the **human-tlink** will ensure that translation equivalent words denote individuals of the same (natural) gender.[2] The use of a common type system makes it easier to express such relationships, although it is not essential, provided that comparable information is encoded in both monolingual representations.

Somewhat rarer and more complex cases of linking arise when a translation mismatch involves transformations operating on FSs. For example, the equivalence class resulting from translation pairs such as 'English:*furniture* ≈ Spanish:*muebles*' can be represented by establishing a link between a word and its translation in the plural form. In this case the equivalence holds between a basic lexical entry and a lexical entry after rule application. In our approach, this behaviour can be modelled straightforwardly, by instantiating one half of the tlink with the appropriate lexical rule. This is shown in the tlink below where the target-language side (`< TFS >`) inherits from the FS representation for the plural rule whose input is the tlink's target input (`mueble`) and output is the tlink's target output.

```
furniture / mueble :
tlink
< SFS : 0 > = < SFS : 1 >
< TFS > <= plural <>.
```

Note that the lexical/morphological rule for plural formation used in the tlink is needed anyway for use in parsing/generation, so that its use in tlinks does not involve introduction of elements other than those needed in the monolingual grammars. The tlink above can be represented figuratively as shown in Figure 5 where unlabelled arrows indicate token identity between FSs. Since the singular form *mueble* would not unify with the feature structure at the end of the output path `< TFS : 0 >`, a translation of *mueble* as *furniture* would not be generated and another tlink would be required to relate the two words when translating a phrase such as *a piece of furniture*. Such a tlink would relate a lexical item to

[2] As we will see later, the morphological generalisation involved in this example can be captured, making it unnecessary to manually specify both tlinks.

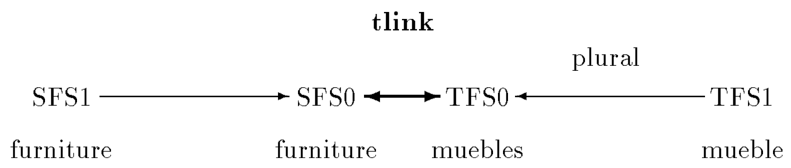

Figure 5: Figurative representation of translation link relating *furniture* and *muebles*

a feature structure representing a partially specified phrase whose full instantiation is established contextually. For example, we can specify that some form of individuating phrase is required rather than providing a literal translation of *mueble* as *a piece of furniture*; this would allow the possibility of *item* as well as *piece*. We can also allow for the possibility of material being interpolated (e.g. *piece of English furniture*). Further details are given in Copestake (1993).

In general, tlinks between a lexical item and a phrase are useful in the treatment of lexical gaps due to a lack of correspondence in sense-extension productivity. For example, a considerable number of verbs can occur as either inchoatives or causatives in English but can only be used as inchoatives in Italian. This phenomenon is particularly recurrent with manner-of-motion verbs, as show in the examples below where the lack of a causative equivalent for verbs such as *fly, run, march* in Italian is resolved by combining inchoative *volare, correre, marciare* with the causative verb *fare*:

(1) a Jamie *flew* his kite
Jamie *fece volare* l'acquilone

b The trainer *ran* the filly at Newmarket
L'allenatore *fece correre* la puledra a Newmarket

c The general *marched* the soldiers
Il Generale *fece marciare* i soldati

The generalization underlying such equivalence can be expressed in terms of diverging lexicalization strategies (Talmy, 1985): English verbs such as *fly, run, march* are allowed to integrate causation, while their Italian equivalents require that causation be expressed separately. Using the tlink mechanism, this divergence can be expressed as an equivalence between a transitive verb and a verb phrase consisting of two verbs (the head and complement daughters) as shown in Figure 6 with the type **tverb-vp-tlink**. The equivalence between the two lexicalization strategies can be semantically expressed by stating that

$$\begin{bmatrix}\textbf{tverb-vp-tlink} & \\ \text{SFS}|0 = \boxed{\textbf{tverb-sign}} & \\ \text{TFS}|0 = \begin{bmatrix}\textbf{vp-sign} \\ \text{DTRS} = \begin{bmatrix}\text{HEAD-DTR} = \boxed{\textbf{vxcomp-sign}} \\ \text{COMP-DTRS} = \langle \boxed{\textbf{iverb-sign}} \rangle\end{bmatrix}\end{bmatrix} & \end{bmatrix}$$

Figure 6: Tlink between a transitive verb (**tverb-sign**) and a verb phrase (**vp-sign**) consisting of a matrix predicate (**vxcomp-sign**) and its verbal complement (**iverb-sign**)

- an English verb describing caused motion such as *fly* corresponds to Italian causative *fare* plus the inchoative Italian equivalent of the English motion verb (*volare*), and
- the agent and theme roles across the source and target sides of the tlink are identical.

In Figure 7, this is indicated in the type **tverb-vp-caused_motion-tlink** where the type **tverb-vp-tlink** (the parent tlink) is augmented with specifications concerning verb semantics. The role types **cause** and **move-manner** which characterize the thematic functionality of agent and theme participants restrict the application of the tlink to verbs expressing caused motion with manner expressed; the identity of the two roles across the source and target sides of the tlink (encoded by the integer tags $\boxed{2}$, $\boxed{3}$) specifies the equivalence in lexicalization strategies.

In some cases the existence of a tlink between two lexical items implies a further translation relationship. For example, in English, there is a regular sense extension such that a word used primarily to denote an animal can also be used to denote the meat of that animal (e.g. *lamb*, *chicken*, see Copestake and Briscoe (1992)). A similar sense extension rule applies to Italian (Östling 1991) but in Dutch a compound is generally used (*lam*, *lamvlees*). The correspondence between the two processes (e.g. sense extension in English and compounding in Dutch) and their import on translation equivalence can be expressed by means of a *tlink-rule* which allows transformations to be defined for tlinks. For example, we can automatically generate the relationship between the FS representing the animal sense of *lamb* (`lamb_1`) and `lamvlees` by transforming the simple tlink between the FS representing the meat sense of *lamb* (`lamb_2`) and `lam` into a new tlink where the source input FS results from applying the sense extension rule 'animal-grinding' to `lamb_1` and the source input FS results from compounding *lam* with *vlees*, as shown diagrammatically in Figure 8. Since we are just making use of the monolingual sense extension mechanism here we can rely on that to handle cases where the sense extension is blocked (e.g. *pig*). It does not necessarily matter for translation purposes whether the lexical rule can fully predict the effects of the sense extension; even if it is used to encode

$$
\left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{tverb-vp-caused\_motion-tlink}} \\
\mathrm{SFS|0} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{iverb-sign}} \\
\mathrm{CAT} = & \boxed{\textbf{ivcat}} \\
\mathrm{SEM} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{tvsem}} \\
\mathrm{PRED} = & \textbf{and} \\
\mathrm{ARG1} = & \boxed{\textbf{vformula}} \\
\mathrm{ARG2} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{binformula}} \\
\mathrm{PRED} = & \textbf{and} \\
\mathrm{ARG1} = & \boxed{2}\left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{thformula}} \\
\mathrm{PRED} = & \textbf{cause} \\
\mathrm{ARG1} = & \boxed{1} \\
\mathrm{ARG2} = & \textbf{obj}
\end{array}\right] \\
\mathrm{ARG2} = & \boxed{3}\left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{thformula}} \\
\mathrm{PRED} = & \textbf{move-manner} \\
\mathrm{ARG1} = & \boxed{1} \\
\mathrm{ARG2} = & \textbf{obj}
\end{array}\right]
\end{array}\right]
\end{array}\right]
\end{array}\right] \\
\mathrm{TFS|0} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{vp-sign}} \\
\mathrm{ORTH} = & \langle \textbf{fare}, \boxed{5} \rangle \\
\mathrm{CAT} = & \boxed{\textbf{vpcat}} \\
\mathrm{SEM} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{vxcompsem}} \\
\mathrm{PRED} = & \textbf{and} \\
\mathrm{ARG1} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{vxcompformula}} \\
\mathrm{PRED} = & \textbf{fare} \\
\mathrm{ARG1} = & \textbf{eve} \\
\mathrm{ARG1} = & \boxed{1}
\end{array}\right] \\
\mathrm{ARG2} = & \left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{binformula}} \\
\mathrm{PRED} = & \textbf{and} \\
\mathrm{ARG1} = & \boxed{2} \\
\mathrm{ARG2} = & \boxed{4}
\end{array}\right]
\end{array}\right] \\
\mathrm{DTRS|COMP\text{-}DTRS} = & \left[\begin{array}{ll}
\mathrm{ORTH} = & \langle \boxed{5} \rangle \\
\mathrm{SEM} = & \boxed{4}\left[\begin{array}{ll}
\multicolumn{2}{l}{\textbf{ivsem}} \\
\mathrm{PRED} = & \textbf{and} \\
\mathrm{ARG1} = & \boxed{\textbf{vformula}} \\
\mathrm{ARG2} = & \boxed{3}
\end{array}\right]
\end{array}\right]
\end{array}\right]
\end{array}\right]
$$

Figure 7: Tlink for equivalences such as 'English:*fly* ≈ Italian:*fare volare*'. The 'cause' (agent) and 'move-manner' (theme) roles of the source-language verb correspond to the 'cause' role of *fare* and the 'move-manner' role of the other target language verb.

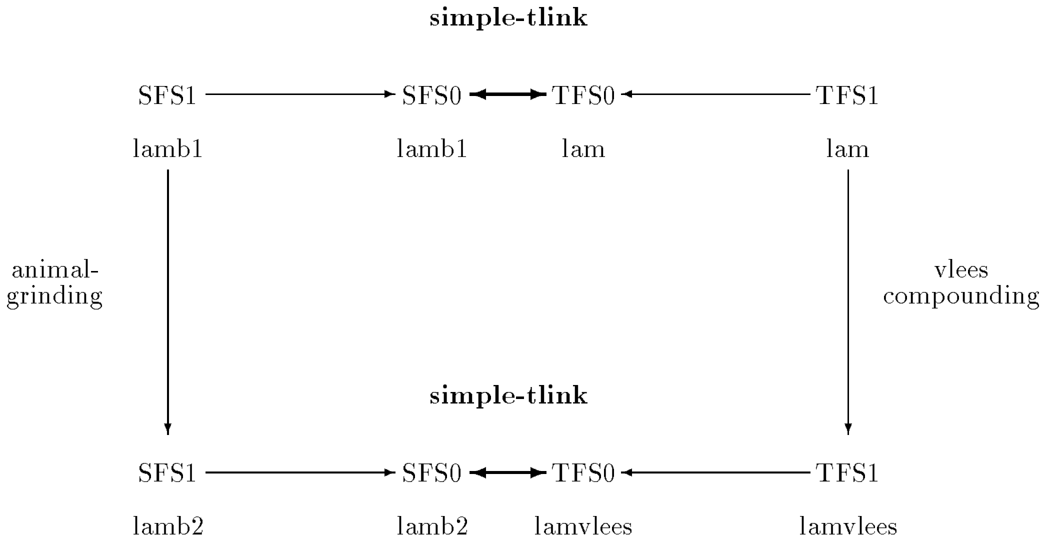

Figure 8: Tlink rule relating animal and meat senses of *lamb* to *lam* and *lamb-vlees*

the regular aspects of the relationship between two existing lexicalised items, an appropriate translation link will be generated if the monolingual processes are sufficiently similar.

There are many examples of such correspondences; for example the English sense extension between trees and their fruits (*pear* etc) is mirrored in Italian with a gender distinction; the trees are masculine but the fruits feminine (*pero*, *pera*). In fact, our earlier examples of *muebles* ≈ *furniture* and *maestro/a* ≈ *teacher* can also be described more generally. A tlink-rule can be defined for human-denoting nouns so that given the equivalence between two entries of one gender the tlink for the other gender can be produced. Since gender in English is normally unmarked, one instantiation of this is with a 'null lexical rule', which states equality between two FSs. Thus given the tlink between *teacher* and *maestra* we can automatically generate that between *teacher* and *maestro* (see Figure 9). The count-mass discrepancy which is responsible for the need to invoke pluralisation in the *muebles* ≈ *furniture* example can also be generalised. Rather than manually specifying the translations of both the singular and plural count nouns for every case we can specify that the translation equivalent of the singular noun will be some more complex construction which individuates the

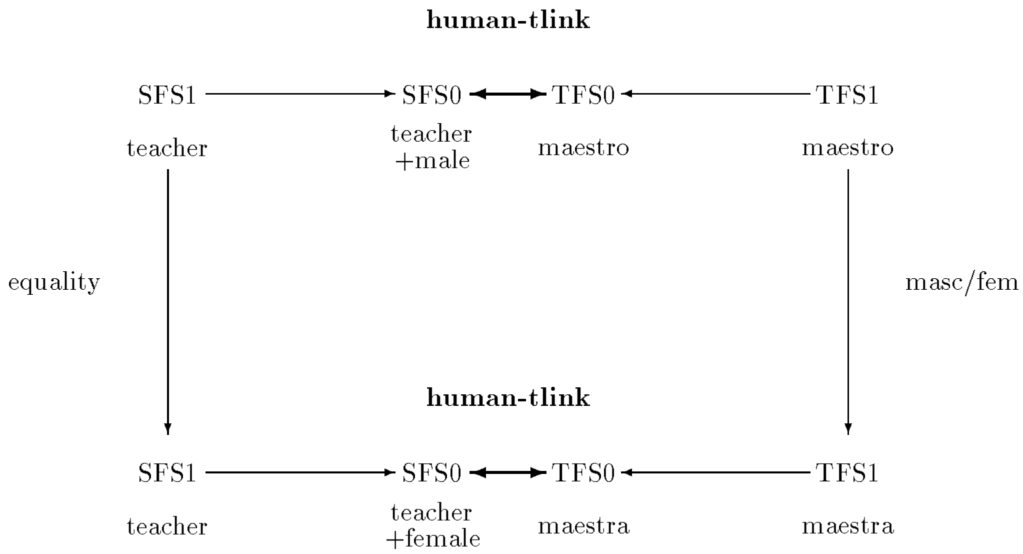

Figure 9: Tlink rule for *teacher* $\approx$ *maestro/a*

mass noun.

As the examples above have shown, multilingual representation requires access to detailed lexical semantic information. Such a requirement does not imply provision of task specific information, since access to detailed lexical semantics is also needed to achieve adequacy in monolingual representation (see, for example, Briscoe *et al.*, 1990; Copestake and Briscoe, 1992). Furthermore, similarity in lexical semantic representation can be used to allow sense selection when constructing tlinks semi-automatically from bilingual dictionaries (Copestake *et al.*, 1992). For example, given that a bilingual dictionary gives *cioccolata*, *cioccolato* and *cioccolatino* as the Italian translations of *chocolate*, we can determine that `chocolate_L_1_4` translates as *cioccolata* because of the comparatively close similarity between their lexical semantic structures.

5 Tlinks in machine translation

General constraint resolution is not a realistic approach to MT because of its computational intractability, but we can classify MT systems in terms of such a model according to the sort of translation constraints they assume, and how they control the process of translation. For example in SRI's BCI system (Alshawi *et al.*, 1991) the source language string is parsed and transfer is carried out on the (quasi-)logical form representation. This produces a quasi-logical form appropriate for the TL, which can be used to drive a head-driven generator. In contrast, the Shake-and-Bake approach (Whitelock, 1992; Beaven, 1992) relies on lexical transfer operating on lexical signs which have had their variable instantiated as a result of parsing. As in the BCI system, transfer operates before generation, but in Shake-and-Bake generation is constrained by 'bags' of instantiated lexical entries rather than an ordered representation of the sentence as a whole. This has the advantage that there is no problem of potential mismatch between the monolingual grammars. However it also means that normal generation techniques are not applicable; the algorithm actually used roughly involves producing all possible orderings of transferred target language signs and using standard parsing techniques to determine which of these orderings meet the constraints of the target language grammar. The ELU system (Estival *et al.*, 1990) demonstrates a third approach; transfer operates at both a lexical and a phrasal level, and transfer and generation are interwoven by making use of transfer variables which indicate explicitly that parts of the structure remain untranslated.

Tlinks contain enough information to potentially be usable to derive the lexical cross-linguistic component in any of these systems, because they all make use of declarative representation techniques and maintain a distinction between the monolingual grammars and the translation relationships. To make use of tlinks in the BCI system, for example, syntactic information would be ignored, for the ELU system, uninstantiated portions of phrasal signs would be explicitly marked by transfer variables. Tlinks could straightforwardly be used to drive the strongly lexicalist approach adopted by Shake-and-Bake, where all correspondences are stated at the lexical level. Because tlinks contain very little additional information beyond that found in the monolingual grammars and lexicons, we do not foresee that the difficulties of exploiting the multilingual LKB in other systems are significantly greater than those involved in the monolingual case.

References

Alshawi, H., D. Carter, M. Rayner and B. Gambäck (1991) 'Translation by quasi logical form transfer', *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91),* Berkeley, California, pp. 161–168.

Beaven, J. L. (1992) *Lexicalist unification-based machine translation,* PhD thesis, University of Edinburgh.

Briscoe, E. J., A. Copestake and B. Boguraev (1990) 'Enjoy the paper: lexical semantics via lexicology', *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90),* Helsinki, pp. 42–47.

Briscoe, E.J., A. Copestake and V. de Paiva (in press) *Inheritance, defaults and the lexicon,* Cambridge University Press, Cambridge, UK.

Copestake, A. (1992) 'The ACQUILEX LKB: representation issues in semi-automatic acquisition of large lexicons', *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92),* Trento, Italy, pp. 88–96.

Copestake, A. (1993) *Constraints, tlinks and MT,* ACQUILEX Working paper.

Copestake, A. and E. J. Briscoe (1992) 'Lexical operations in a unification based framework' in J. Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation. Proceedings of the first SIGLEX Workshop, Berkeley, CA,* Springer-Verlag, Berlin, pp. 101–119.

Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni and E. Marinai (1992) 'Multilingual lexical representation' in A. Sanfilippo (eds.), *The (other) Cambridge ACQUILEX papers,* University of Cambridge Computer Laboratory. Technical Report No. 253, pp. 117–129.

Estival, D., A. Ballim, G. Russell and S. Warwick (1990) 'A syntax and semantics for feature structure transfer', *Proceedings of the 3rd International Conference on theoretical and methodological issues in MT of NLs (TMI-90),* Austin,Texas, pp. 131–143.

Kaplan, R., K. Netter, J. Wedekind and A. Zaenen (1989) 'Translation by Structural Correspondences', *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics (EACL-89),* Manchester, UK, pp. 272–281.

Östling, A. (1991) *Sense extensions in the Italian food subset,* ACQUILEX working paper, Dipartimento di Linguistica, Università di Pisa.

Procter, P. (ed) (1978) *Longman Dictionary of Contemporary English,* Longman, London.

Pustejovsky, J. (1991) 'The generative lexicon', *Computational Linguistics, vol.17(4),* 409–441.

Sanfilippo, A. (in press) 'NLP Encoding of Lexical Knowledge' in Briscoe, E.J., A. Copestake and V. de Paiva (eds.), *Inheritance, defaults and the lexicon,* Cambridge University Press, Cambridge, UK.

Sanfilippo, A., E. J. Briscoe, A. Copestake, M. A. Marti and A. Alonge (1992) 'Translation equivalence and lexicalization in the ACQUILEX LKB', *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92),* Montreal, Canada.

Sanfilippo, A. and V. Poznanski (1992) 'The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources', *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92),* Trento, Italy, pp. 80–88.

Shieber, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar,* CSLI Lecture Notes 4, Stanford CA.

Talmy, L. (1985) 'Lexicalization Patterns: Semantic Structure in Lexical Form' in T. Shopen (eds.), *Language Typology and Syntactic Description 3. Grammatical Categories and the Lexicon,* Cambridge University Press.

Whitelock, P. (1992) 'Shake-and-bake translation', *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92),* Nantes, France.

Zajac, R. (1989) 'A transfer model using a typed feature structure rewriting system with inheritance', *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL-89),* Vancouver, BC, pp. 1–7.