, , 1–35 ()


# Acquisition of lexical translation relations from MRDs

ANN COPESTAKE, TED BRISCOE aac@csli.stanford.edu
*University of Cambridge, Computer Laboratory*

PIEK VOSSEN piek.vossen@let.uva.nl
*Computer Centrum Letteren, Universiteit v. Amsterdam*

ALICIA AGENO, IRENE CASTELLON, FRANCESC RIBAS,
GERMAN RIGAU, HORACIO RODRÍGUEZ, ANNA SAMIOTOU horacio@lsi.upc.es
*Facultad de Informatica, Universitat Politècnica de Catalunya*

*Received September, 1994; Revised July, 1995*

**Abstract.** In this paper we present a methodology for extracting information about lexical translation equivalences from the machine readable versions of conventional dictionaries (MRDs), and describe a series of experiments on semi-automatic construction of a linked multilingual lexical knowledge base for English, Dutch, and Spanish. We discuss the advantages and limitations of using MRDs that this has revealed, and some strategies we have developed to cover gaps where no direct translation can be found.

**Keywords:** Machine readable dictionaries, lexical acquisition, lexical knowledge bases

## 1. Introduction

The research reported here is part of the ACQUILEX II project which has as one aim the automatic or semi-automatic construction of fragments of a multi-lingual lexical knowledge base containing detailed syntactic and semantic information from MRD resources. It has been estimated that the average time needed to construct a lexical entry for a natural language processing system manually is about 30 minutes [32]: since a lexicon must contain at least 40,000 words to achieve even moderately wide coverage [53] it is clear that techniques which either partially or totally automate this process should be investigated.

Three major classes of techniques for lexicon acquisition have been developed: machine-aided manual construction, extraction from corpora and extraction from MRDs. Each of these techniques has advantages and disadvantages. Various tools have been developed which make manual lexicon construction and maintenance quicker and easier, and to some extent avoid the need for detailed linguistic knowledge. This approach has been taken in large-scale MT systems, see for example, [34] and the description of METAL in [21]. Manual construction is the most reliable technique, but it is time-consuming and errors of omission, in particular, are a problem. Automatic or semi-automatic extraction of information from corpora is very promising, but a vast amount of data is needed for reliable entries to be constructed on any but the most common words, and this is currently not easy to obtain for most languages. Unless the corpus closely matches the intended text type for the NLP system, relevant senses of words are likely to be missing. Furthermore, extraction of lexical semantic information from corpora has not yet been accomplished on a large scale. The techniques we describe here rely on the use of MRDs. These can be time-consuming to process initially, may contain errors and inconsistencies, and are of highly variable quality with respect to their utilisation for NLP. But they do offer broad coverage and it is possible to extract quite extensive lexical semantic information. MRDs are particularly suitable for providing a labour-saving resource to augment a manually constructed core lexicon (for further discussion see [7] and [8]).

Earlier work within the ACQUILEX project concentrated on building monolingual lexical fragments, both to investigate their potential utility for tasks that do not involve translation, and as a basis for constructing the multilingual lexical knowledge base (LKB). Large scale monolingual lexicon fragments have been constructed semi-automatically for four languages (English, Spanish, Dutch and Italian) from MRDs; see, for example, [13], [41], [49] and [1]. Here we are concerned with the use of the LKB to represent multilingual information in the form of links between monolingual lexical entries, which we refer to as *tlinks* (translation links), and with the methodology for constructing tlinks from the available MRD resources.[1]

Although there is a considerable amount of published work on the use of MRDs to extract monolingual lexical entries, far less attention has been paid to MT lexicons. Helmreich *et al.*[20] report on the use of MRDs in the Pangloss project, but this essentially involved manual construction of lexical entries, aided by a sophisticated user-interface to the MRDs, although procedures to automate acquisition were being developed. Neff and McCord [32] describe the direct and indirect exploitation of multiple MRDs to support the English-to-German version of the Slot Grammar based machine translation system, LMT. This builds on the extensive work carried out to construct a monolingual English lexicon (see [24]), using information extracted from the bilingual Collins English-German dictionary to construct transfer relations. However, Neff *et al.*[31] report that the automatically generated entries often have to be revised by hand. It seems to be generally agreed that bilingual MRDs alone are insufficient for constructing a lexicon for MT and so in the work reported here we have followed Neff *et al.* in combining the use of monolingual and bilingual MRDs. Our aims are somewhat different from theirs, however, because we are attempting to construct a lexical knowledge base that will support bidirectional translation between three languages in a range of formalisms, rather than a unidirectional pairing for a single system. We also place more emphasis on detailed lexical semantic information.

The general strategy we have adopted is to relate lexical entries corresponding to word senses in Spanish and Dutch to the English entries, using an approach which combines the use of bilingual MRDs with comparison of syntactic and semantic aspects of the lexical entries constructed using the monolingual dictionaries.

English was chosen as the central language because of the relative richness of the monolingual lexical entries, which was largely due to the use of the Longman Dictionary of Contemporary English (LDOCE [35]) and the extensive work on analysis of LDOCE that has been carried out: see, in particular, [55], [11], [52], [49]. Since it is more difficult to extract information from the available MRDs for other languages, one of the aims of constructing a multilingual LKB is to augment the other lexicons by transfer of English information.

In general, our use of MRDs is aimed at extraction of sufficient information in order to categorise lexical entries according to predefined linguistically motivated classes. Thus, for example, verbs can be distinguished not only syntactically as intransitive, transitive and so on, but also into semantic classes such as psychological verbs, verbs of perception and so on (see, for example, [26]). Sanfilippo and Poznanski [41] demonstrated how MRDs could be used to semi-automatically classify verbs into such classes, which correspond to types in the typed feature structure based lexical representation language (LRL) used in the LKB system. Our methodology for automatically classifying nouns semantically is based on the extraction of taxonomies from MRDs. For example, the definition of *Sauternes* in LDOCE is:

**Sauternes** a type of sweet gold-coloured French wine

It is possible to parse this definition to extract the *genus term*, in this case *wine*, and then to disambiguate this with respect to the senses given for *wine* in LDOCE, giving *wine*$^1$ 1 (the superscript refers to the homonym number and the following numeral to the sense within the entry). This sense of wine has the disambiguated genus term *drink*$^2$ 1. For details of extraction and disambiguation of genus terms, see [25], [47], [12], [1].

If we extract genus terms in this way we can derive taxonomies, which are (basically tree-structured) hierarchies of word senses. We can use these to classify nouns with respect to semantic classes such as *comestible substance*: all senses under *drink*$^2$ 1 in the taxonomy will belong to this class, for example. As with verbs, such classes can be motivated by considering derivational morphology and systematic polysemies or sense extensions. For example, most mass terms denoting types of drink can be *portioned* into count nouns denoting a (conventional) portion of that drink: *beer/a beer* etc.

As discussed in detail by Vossen and Copestake [51], there are difficulties in constructing and interpreting dictionary taxonomies, and the classification cannot be carried out totally automatically. It is also comparatively more difficult to extract taxonomies from dictionaries which do not have the controlled vocabulary and semantic coding found in LDOCE. However, most concrete nouns can be straightforwardly located in taxonomies. Taxonomies are also used to establish inheritance relationships between lexical entries represented within the LKB. Although extraction of the genus term is the most important step in classifying a sense, and thus establishing its semantic type, the other parts of the definition, the *differentia*, can also be parsed and provide information which refines the classification.

The representation language used in the LKB, the LRL, is designed to be quite general, capable of encoding a variety of linguistic approaches, following the same

$$\begin{bmatrix} \textbf{lex-noun-sign} & \\ \text{ORTH} = \textbf{"Sauternes"} & \\ \text{CAT} = \textbf{noun-cat} & \\ \text{SEM} = & \begin{bmatrix} \textbf{unary-formula-entity-arg1} \\ \text{PRED} = \textbf{sauternes\_L\_0\_0} \\ \text{ARG1} = \textbf{entity} \end{bmatrix} \\ \text{RQS} = & \begin{bmatrix} \textbf{c\_art\_subst} & \\ \text{ORIGIN-AREA} = \textbf{french} & \\ \text{QUAL} = & \begin{bmatrix} \text{COLOUR\_SPEC} = \begin{bmatrix} \text{COLOUR} = \textbf{gold} \end{bmatrix} \\ \text{TASTE} = \textbf{sweet} \end{bmatrix} \\ \text{PHYSICAL\_STATE} = \textbf{liquid\_a} & \\ \text{FORM} = & \begin{bmatrix} \text{SHAPE} = \textbf{non-individuated} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

*Figure 1.* Simplified lexical entry for *Sauternes*

sort of philosophy as PATR-II [42] and, like it, based on the use of feature structures (FS) and unification. The LRL has a relatively rich description language, which is geared towards lexical representation since it allows for inheritance and default inheritance between lexical entries. The underlying language utilises typed feature structures [10], and can be used for either rule-based or constraint-based approaches to parsing and generation. The extra expressiveness that types provide compared to untyped feature structures is an important component of our approach to lexical representation. For ACQUILEX we adopted a common type system to encode properties of lexical items in the four languages investigated within the project. This ensures compatibility of representation without ruling out language specific parameterisation where necessary. A lexical entry for *Sauternes*, extracted from LDOCE, is partially shown in Figure 1. The types (shown in bold font) are mostly intended to be language independent with the exception of strings (e.g. the value of ORTH) and predicates with appended sense identifiers (e.g. **sauternes_L_0_0** — the 'L' identifies the sense as coming from LDOCE; in general, the first integer stands for the homonym number and the second for the sense number, but here the 0_0 indicates that no sense distinctions were made for this entry).

One important feature of our approach is our emphasis on representing detailed lexical semantic information. We refer to this as *relativised qualia structure* or RQS [9]. The features found in an RQS vary according to the type of the concept being represented; thus COLOUR is an appropriate feature for physical objects and substances (including liquids like *Sauternes*), while TASTE is only relevant for commonly ingested things. Some features in the RQS are instantiated by concepts which are intended to be language independent (e.g. **sweet**) while others are language dependent.[2]

Although the RQS framework was originally described [9] as being related to the concept of qualia structure due to Pustejovsky (e.g. [36]), it is important not to conflate the two: the RQS type system has been developed to a large extent on the basis of analysis of the definitions in dictionaries and is motivated more by empirical data than theoretical considerations, while Pustejovsky's qualia structure is part of his theory of the Generative Lexicon. There is, of course, considerable debate about whether the sort of information we represent in the RQS can properly be regarded as lexical or whether it is part of real world knowledge and indeed, whether any distinction can be drawn between the two. However, it is not clear that this distinction is of great practical significance in our case: we include the RQS information in what we refer to as a lexical KB, but maintain a clear separation between it and the more conventional lexical information. Furthermore, although in a full MT system it is clear that logical inference mechanisms that are not available in the LKB would be required to act on the conceptual information, this is not a requirement for our current work. In fact, MRDs can only supply a rather limited amount of real world knowledge of the sort required to make complex inferences. But they do allow word senses (or concepts) to be hierarchically related to each other and also to be classified with respect to a fixed set of semantic properties, and the type based formalism of the LKB supports this. The LRL is actually somewhat less complex than many variants of the typed feature structure approach advocated for purely linguistic representation purposes, and the underlying language has not been extended for the representation of the RQS.

We can regard the RQS as locating a concept in a multi-dimensional semantic space, according to the set of features appropriate for its type. We cannot hope to have a complete account of meaning of a concept in this manner, but what is required here is categorisation, not complete decomposition: to the extent that we make use of a language independent set of types and features, the RQS thus provides a partial interlingua. The problem of constructing a multilingual lexicon can be seen as relating concepts which can be expressed lexically in either the source or target languages or both. In the simplest cases, translation equivalence can be represented as a one-to-one mapping between senses in the monolingual lexicons. We will assume for the moment that strictly translation equivalent senses should have compatible RQS, though as we will see later this is not always achieved in practice. In any case we do not think of translation equivalence as identity of concepts: there will inevitably be some difference in denotation, except in the rare cases of concepts with an accepted expert definition, such as chemical elements. However speakers of the same language are unlikely to agree precisely on the meaning of a term. So, accepting this indeterminacy, we can take strict translation equivalence as implying that the difference in meaning is not significantly greater than the monolingual spread. For a complete lexicon it would also be necessary to take into account discourse function, conventional implicature and so on, but we will ignore this here.

In theory, it might be possible to simply link lexical entries by comparing their RQSs without the aid of a bilingual dictionary.[3] Unfortunately we have found that it is not possible to accurately construct translation equivalence pairings simply by comparing the RQSs of structures in the monolingual lexicons, because the information we can extract from the MRDs does not give a fine enough grain of discrimination and, in many cases, there is no straightforward one-to-one mapping between lexical entries. We therefore have to make use of bilingual dictionaries in constructing mappings. This is discussed in detail in §3, but first we describe the translation link formalism.

## 2. Translation links

Several unification based formalisms have been proposed 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. [22], [56], [19], [4]). In designing the translation link formalism we have attempted to maintain these advantages, but to abstract away from the aspects of these systems which are specific to particular MT techniques or details of the grammar formalism adopted. We have attempted to maximize the functionality of the system with respect to the expression of linguistic and lexicographic generalisations in order to produce an LKB which would support a variety of approaches to MT, although it is naturally most appropriate to the more lexicalist frameworks.

We define lexical translation equivalence in terms of cross-linguistic links, *tlinks*, stated in terms of lexical entries in the monolingual lexicons and of classes of lexical entries. 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 means that the description of translation equivalence is as compact as possible and ensures that the multilingual and monolingual components are compatible. The use of the typed feature structure formalism makes it easy to define classes of translation equivalence appropriate for particular classes of mismatch.

The full translation link formalism is described in [14], which also discusses how translation links can be interpreted as constraints which must hold between source and target language structures in order for them to be translation equivalents. The formalism can be used as the basis for a constraint based transfer system, although for practical use it would be necessary to pay much more attention to control aspects. Some complex translation mismatch cases are described in [40] and [17]. Here we will concentrate on informally describing the simpler classes of tlinks, since these make up the vast majority of cases, including all those we can currently extract semi-automatically from MRDs. However at the end of the section we give a sketch of a more complex head-switching example. We begin with an example of a tlink for English *wine* ≈ Spanish *vino*, which is written as follows in the description language:

```
wine_1 / vino_1 :
simple-tlink.
```

This tlink directly relates monolingual lexical entries and expands into a typed FS which includes both lexical signs. If we assume for the moment that this is the only English-Spanish tlink for *wine*, it describes a constraint which states that any English language parsed structure which contains the lexical sign for this sense of *wine* has to be paired with a Spanish structure containing the sign for *vino*, with the same variable being used in the semantics of both signs. This equivalence of variables can be used to drive translation, as in Shake'n Bake transfer [54], although the implementation used in the LKB is somewhat different. The form that the tlink takes in terms of typed feature structures is determined by its type, **simple-tlink**, which is a subtype of **tlink**.

Although this tlink relates lexical entries directly, the formalism has to allow for greater flexibility. All tlinks are defined as FSs of type **tlink**, which contains the source FS (path < SFS >) and the target FS (< TFS >) which are both of type **rule**. This is the type which is used to encode monolingual lexical rules and grammar rules: minimally, a rule establishes a correspondence between an input sign (1) and an output sign (0). The rule-inputs of tlinks (< SFS 1 > and < TFS 1 >) are meant to be instantiated by unification with lexical entries for word senses in the source and target languages; the rule-outputs (< SFS 0 > and < TFS 0 >) provide the translation equivalent structures when this is done.[4] It is the pairing between the rule outputs which is used during translation itself: tlinks are applied by unification of their source language output structure with an instantiated sign in the FS representing the parse of the source sentence, giving a target language output sign which must correspond to part of the target language structure. Thus tlink application creates a set of constraints which must be satisfied by the target language structure and which can be resolved to generate the target language sentence. However, the explicit connection to the monolingual lexical entries allows the description of tlinks to be compact and consistent with the monolingual lexicon, while being sufficiently expressive to allow for translation mismatches.

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, such as the example given above, can be represented as **simple-tlink**s. A **simple-tlink** is applicable in the case where two lexical entries which denote one-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. The particular paths equated will depend on the way that the type system is used to encode semantic structure. Rather than showing the FS expansions of tlinks we will use a diagrammatic representation as in Figure 2 which shows the **simple-tlink** for English:*wine* $\approx$ Spanish:*vino*. The formal semantic structure corresponding to each sign is shown in the figure using a linearised form of the representation (the semantic theory used is Indexed Language, InL, [57] — the actual encoding uses FSs but this notation is considerably easier to read). The coindexation between the

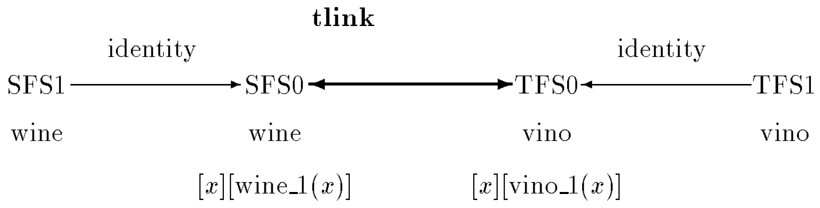


*Figure 2.* Figurative representation of translation link relating *wine* and *vino*

arguments of the predicates of the source and target language is indicated by the use of the same variables: in this case, no other information is shared.

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 : 1 > = human
< TFS : 1 > = human
< SFS : 0 : RQS : SEX > = < TFS : 0 : RQS : 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 `< RQS : 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.[5]

Somewhat rarer and more complex cases of linking arise when a translation mismatch involves operations 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.

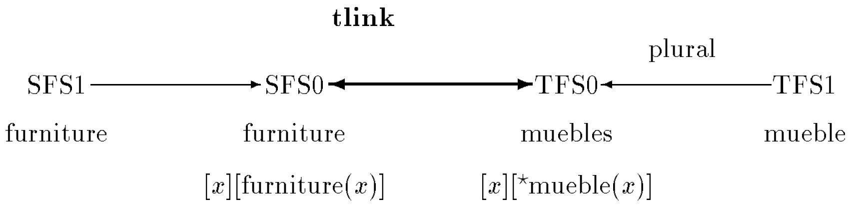


*Figure 3.* Figurative representation of translation link relating *furniture* and *muebles*

```
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 can be represented figuratively as shown in Figure 3 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*. The tlink mechanism allows this to be done in a variety of ways. The exact phrase could be encoded:

```
piece_1 of_1 furniture_1 / mueble_1 :
phrasal-source-tlink .
```

where the type **phrasal-source-tlink** specifies that the English words are combined by ordinary grammar rules to produce that exact phrase. In general, we want to weaken this to allow for the possibility of material being interpolated (e.g. *piece of English furniture*): we do this essentially by establishing a semantic rather than a syntactic connection between the lexical signs (see [14] for details). Furthermore we would like to allow for a range of individuating words: *item*, *article* and even *stick* are possible instead of *piece*. Such a tlink would relate a lexical item to a feature structure representing a partially specified phrase whose full instantiation is established contextually, since the determination of the particular individuating word is properly part of the monolingual rather than the bilingual lexicon. The formalism allows for underspecified placeholders for lexical items: for example

```
individuator of_1 furniture_1 / mueble_1 :
phrasal-source-tlink .
```

where `individuator` is a FS which subsumes the FSs for *piece*, *item*, etc. Of course, we could generalise over cases of this sort and define a tlink type which incorporates the lexical sign for *of* rather than mentioning it explicitly each time.

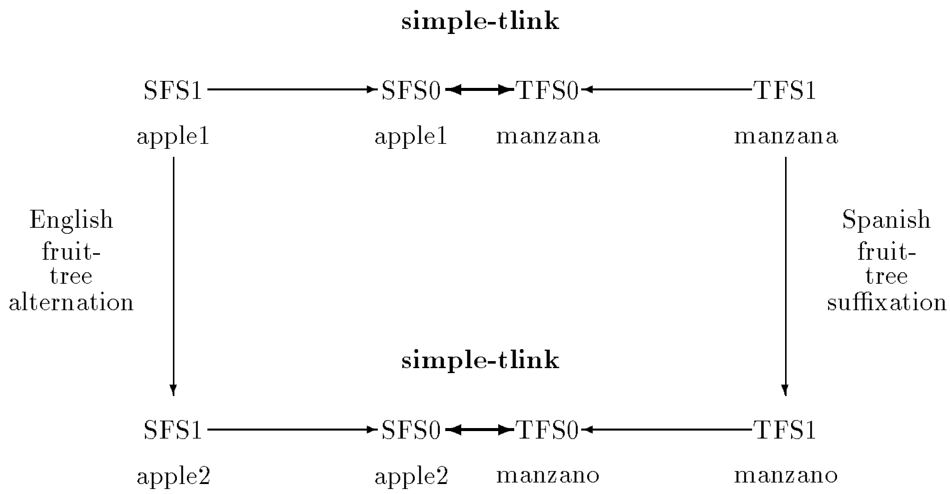


*Figure 4.* Tlink rule relating fruit and tree senses of *apple* to *manzana/o*

We also make use of *tlink-rules*, which are FSs which can be regarded either as relating tlinks, or as relating monolingual lexical or morphological rules. For example, the portioning sense extension mentioned in the introduction is found both in English and in Dutch, so that if we have a tlink relating the mass sense of *gin* to *jenever* we can automatically generate the relationship between the count senses of both terms.[6] More interesting cases involve some mismatch between languages, so for example, while in English nouns denoting fruits can also normally be applied to the tree or plant which produces them (*apple*, *apricot*, etc.), in Spanish the same polysemy normally involves a change in gender or suffixation (*manzana*, *manzano*; *albaricoque*, *albaricoquero*). A tlink-rule expressing this relationship is shown diagrammatically in Figure 4. Sometimes affixation in Spanish seems to correspond to compounding in English: for example, the suffix *-al* applies to plants to give a sense denoting a group of such plants, that is a field or plantation, etc. For example, *aguacate* (*avocado*) takes the suffix to form *aguacatal* (*avocado plantation*). Marti and Soler [27] discuss these examples and a number of others which have been found by analysis of the bilingual dictionary Vox-Harrap's Esencial [6]. This is an important issue for automatic acquisition of tlinks, because it indicates that some translation equivalences can be systematically derived from others.

Finally, we briefly consider a head switching example, which further demonstrates the possibility of automatically deriving complex relationships from simple ones and also illustrates that the tlink mechanism is adequate to deal with one of the standard test cases for MT formalisms. English verbs of manner of motion (e.g. *swim*, *stagger*) can occur with a locative expression describing a completed path. In many other languages, including the Romance languages and Japanese, this pattern is not possible [45]. Here we will concentrate on Spanish, where the manner of motion is expressed with an adverbial, while the main verb conveys the path. For example, (1a) is usually translated as (1b) (which is literally translated as (1c)).

(1) a Kim swam across the river.
    b Kim cruzó el río nadando.
    c Kim crossed the river swimming.

The use of tlinks to treat this example was discussed in detail in [40]. Some details of the tlink formalism have changed, but we will follow the same semantic analysis here. The verb semantic representation uses a neo-Davidsonian approach with proto-roles [18] being used to express thematic relations. The semantics is described in detail in [39]. We assume that strict intransitive motion verbs such as *swim* are related by lexical rule to entries subcategorized for a PP expressing a path. The details of this are unimportant here, but we assume that in Spanish this rule cannot apply to manner of motion verbs when the path is bounded. The logical form for (1a), expressed in a linearised notation, is:

$$\exists e, y[\mathrm{swim}(e) \wedge \text{p-agt-cause-move-manner-path}(e, \mathrm{kim}) \wedge \mathrm{across}(e, y) \wedge \mathrm{river}(y)] \quad (1)$$

We assume that the gerundive in the Spanish sentence expresses a distinct event, which is related to the main verb event by a predicate **while** indicating that the first event is temporally contained within the second. Thus (1b) has the following semantics:

$$\begin{aligned}\exists e, e', y[&\mathrm{cruzar}(e) \wedge \text{p-agt-cause-move-path}(e, \mathrm{kim}) \wedge \text{p-pat-path}(e, y) \\ &\wedge \mathrm{while}(e', e) \wedge \mathrm{nadar}(e') \wedge \text{p-agt-cause-move-manner}(e', \mathrm{kim}) \wedge \text{río}(y)]\end{aligned} \quad (2)$$

The translation link which expresses the equivalence has to relate the pair of lexical entries for *swim* (with completed path) and *across* to the pair for *cruzar* and *nadando*. As above, we achieve this by relating phrases which are constrained to include FSs corresponding to those lexical items in the relevant semantic relationship, without specifying the syntactic structure directly. The tlink is shown diagramatically in Figure 5 ('incl' indicates the inclusion relationship). The translation of *the river* as *el río* occurs as usual, and the coindexation on the variables ensures that it is expressed as an argument to *across* in the English sentence and to *cruzar* in Spanish.

Clearly it would be very time-consuming to have to construct such relationships manually for every possible pairing of a manner of a movement verb with a preposition expressing a path. Instead, we generate this equivalence automatically, from the straightforward relationship between strict intransitive *swim* and *nadar* plus a relationship between *cruzar* and an underspecified English movement verb taking an *across* PP as a path argument. This is done by a tlink rule which relates the English lexical rule which creates the path-taking manner-of-movement verb sense to the Spanish suffixation with *+ndo* (see [17], [40]).[7]

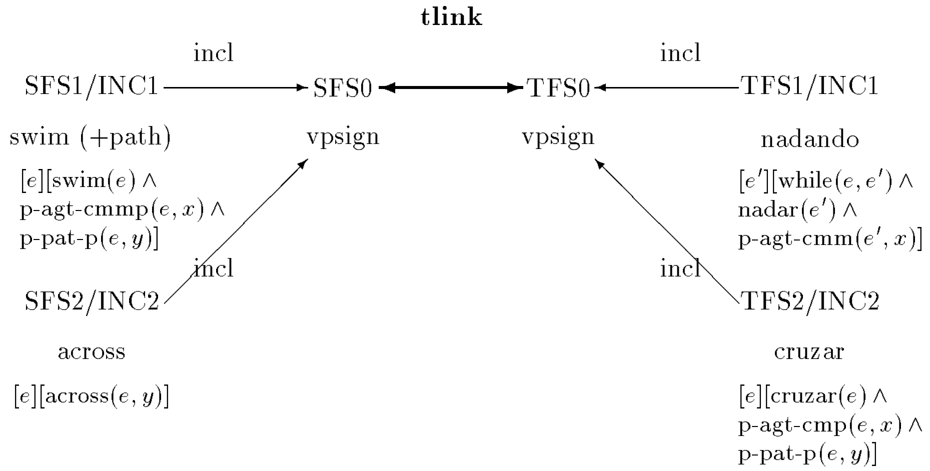


*Figure 5.* Manner of movement with completed path tlink. The following abbreviations are used in the semantic representation: p-agt-cmmp for p-agt-cause-move-manner-path, p-agt-cmm for p-agt-cause-move-manner, p-agt-cmp for p-agt-cause-move-path, and p-pat-p for p-pat-path.

## 3. Constructing tlinks

Since the monolingual ACQUILEX LKB on the whole uses dictionary sense distinctions, the linking between monolingual entries has to make reference to these. The following example illustrates the sort of tlink we aim to derive:

```
wine_L_1_1 / vino_X_1_1 :
simple-tlink.
```

This tlink directly relates lexical entries corresponding to dictionary senses: `vino_X_1_1` stands for the VOX monolingual Spanish dictionary [5] sense *vino*[1] 1, L stands for LDOCE, as before. Since we were only dealing with nouns denoting physical objects, there was no need to attempt to determine complex variable equivalences — thus all the tlinks derived could be assumed to have the same sort of equivalence as a `simple-tlink`. Various other subclasses of tlinks are discussed below: these were defined either to record information about how the tlink was derived or to indicate its partiality — they do not affect the internal structure.

Two sets of experiments were carried out on semi-automatically deriving translation links: one linking Dutch and English lexicons, the other Spanish and English. In both cases, a similar basic approach was taken, which we describe initially, followed by details of the Dutch-English and Spanish-English experiments. The methodology involves looking at each source sense in turn and attempting to find corresponding senses (possibly modified or combined into phrases) in the target language. As we mentioned in the introduction, although in principle it might be possible to relate monolingual lexical entries purely by comparison of their RQSs, in practice the definitions are too coarse grained for this to be possible, especially since lexicographers may choose to highlight different features in their definitions, and thus we have to check for compatibility of structures, rather than an exact match. Furthermore, since many lexical entries only have phrasal equivalents, we have to use bilingual dictionaries to constrain the search space.

The use of bilingual dictionaries means we have to consider the following difficulties:

1. There is no direct correspondence between the senses given in the monolingual dictionaries, the bilingual dictionaries and the translations.
2. A source sense may be missing in the bilingual dictionary.
3. A source sense may be found in the bilingual dictionary but its translation is not found in the target monolingual dictionary.
4. A source sense is found in the bilingual dictionary but it is translated by a phrase.
5. A translation is restricted by a label (such as *med.* for medical).

The sense correspondence problem can be appreciated most easily by looking at an example (taken from [50]):

| Van Dale monolingual | | VanDale Dutch-English bilingual | | | LDOCE | |
|---|---|---|---|---|---|---|
| Entry | Senses | Entry | Senses | Translations | Entry | Senses |
| doos | 01⟨*object*⟩ | doos | 01⟨*object*⟩ | box | box | 01 *container* |
| | | | | | | 02 *the contents* |
| | | | | | | 03 *room* |
| | | | | | | 04 *piece of metal* |
| | | | | | | 05 *television* |
| | 02⟨*toilet*⟩ | | 02⟨*toilet*⟩ | loo (BE) | loo | 01 *lavatory* |
| | (informal) | | (informal) | lav (BE) | GAP | |
| | | | | John (AE) | John | 01 *man* |
| | | | | | | 02 *lavatory* |
| | | | | can (AE) | ... | |
| | GAP | | 03⟨*jail*⟩ | clink (BE) | ... | |
| | | | (informal) | junk (BE) | ... | |
| | | | | the slammer (AE) | ... | |
| | | | | the can (AE) | ... | |

The words in italics are included from the definitions in order to indicate the senses involved, the words in angle brackets are English glosses of the Dutch. We have not listed all the LDOCE senses for *can*, *clink*, *junk*, etc. because there are too many irrelevant ones. It is clear that we have to isolate the appropriate sense of *box*, *can*, etc., in order to produce the correct tlinks and on the whole bilingual dictionaries do not provide an adequate source of information for doing this. This example also illustrates the problem of gaps in the dictionaries: *lav* is not given as a headword in LDOCE, for example. (This could alternatively be regarded as an example of a bilingual dictionary giving an apparently inappropriate translation, since *lav* is not a common word in contemporary British English in our experience.) Finally, it shows examples of phrases used in the translation field: *the slammer* and *the can*. However these are not genuine phrasal translations but instead indications of a restriction on the use of the terms in the monolingual grammar since *the* is obligatory with these senses:

(2) * He was in a can.

We originally attempted to tackle the bilingual sense disambiguation problem by developing a utility program, LUCIFER, for determining the best match among multiple target lexical entries, by comparing them for their similarity to the source. For example, when constructing a tlink for the first sense of *doos* given in the Van-Dale monolingual dictionary, its FS would be compared with the FSs generated from LDOCE for the different senses of *box* and the most similar proposed to the user as the appropriate translation. LUCIFER was designed to learn which features in the sign were the most reliable indicators from user feedback accepting or rejecting the proposed tlinks (see [16]). However this technique has not been exploited to the limit because we adopted the methodology of linking semantically restricted subsets of the monolingual lexicon. It was not feasible to link entire dictionaries given the limited resources of the project so semantic selection of subsets was necessary to ensure that compatible sets of words were used for each language. This preselection greatly reduced the ambiguity problem. For example, English *dish* in the food sense can be translated as *gerecht* in Dutch but the container sense is translated as *schaal*. By preselecting the food subset, we ensure that we consider only the food senses, that is *dish*$^1$ 2 and *gerecht*, but exclude *schaal*. The semantic classification used in determining the subset was derived semi-automatically from the MRDs, primarily from the automatically derived taxonomies mentioned above [47], [12], [1].

The work reported here relates to noun senses denoting foods or drinks. Relatively few words have multiple senses denoting foods or drinks, with the exception of regular polysemy, in particular the portioning sense extension mentioned in the introduction. The type/kind usages are also occasionally given distinct senses and sometimes the two are conflated, as in the LDOCE example below.

**brandy 1** a strong alcoholic drink ...

**brandy 2** a type or single drink of this

Excluding these cases, we were left with very few examples of ambiguity to resolve with LUCIFER because we had, in effect, preempted its functionality. In fact, in the work reported in detail below, the Dutch-English experiment relied entirely on the disambiguating effect of the taxonomies, with manual intervention to resolve the remaining distinctions.

Working from a subset is effectively equivalent to taking the semantic classification as an absolute restriction on classifying lexical entries as translation equivalent. For example, food nouns in English could only be related to food nouns in Spanish. It is clear that this might be overrestrictive, since concepts like *soup*, *broth* and *beef tea* are somewhat borderline between foods and drinks, and one can imagine discrepancies in classification between lexicographers. There will also be some errors in the automatic assignments of semantic types to senses. This will tend to make the results reported here worse than would be expected for a more flexible approach which allowed links to be proposed between senses with different semantic types if no other match could be found.

We now turn to the actual experiments, which although adopting the same basic approach differ considerably in detail, partly because of the nature of the dictionary resources available and the difference between Spanish and Dutch with respect to noun-noun compounds. These sections are summaries of [50], [3], [37] and [38]. Both experiments relied on the same subsets of English noun lexical entries, extracted from LDOCE. For this experiment the subsets were extracted according to taxonomic information alone, for example the drink subset consists of all senses found under the relevant sense of *drink* in the taxonomy. The LDOCE drinks subset contained 192 entries, the food subset 260 entries.

### 3.1. Dutch/English

The monolingual Dutch dictionary used to extract the subset was the Van Dale dictionary of contemporary Dutch [43], the bilingual dictionaries used to construct tlinks were the Van Dale Dutch-English [29] and English-Dutch [28] which we will abbreviate as VDE and VED. The bilingual dictionaries are claimed by Van Dale to be largely compatible with the monolingual dictionary, since they are based on a single database of Dutch. There is a considerable disparity in size between the monolingual VanDale which contains about 61000 noun entries and LDOCE with 24000.

Vossen [50] analysed the translations in the VDE for matches to LDOCE and found that of 51500 simple translation fields for nouns[8], only 14000 corresponded to LDOCE entries. Thus only 27% of VDE translations are found directly in LDOCE and only 60% of LDOCE entries are found as translations in the VDE. Most of the mismatches are phrases of some sort, many being adjectivally modified nouns, or noun-noun compounds of varying degrees of transparency. For example: *dierennummer/animal act*, *dierenverzorging/animal care*, *dierenkliniek/animal clinic*, *dierentaal/animal language*.

The VDE and the VED both give two types of translation: main translations, which are supposed to be applicable in most contexts, and secondary translations, which are more specific or restricted. The latter are always given as alternatives to main translations. For example, *baas* (owner of an animal) is given the main translation *owner* and secondary translations *master* and *mistress*. No distinction was made between main translations and secondary translations for this experiment (but see §4 below). Various labels are also found on translations: for instance, in the translations of *baas*, *master* and *mistress* were labelled m. and f. for masculine and feminine. Vossen [50] discusses this in some detail, but since a large set of labels were used and they were applied somewhat unevenly, they were ignored for this experiment.

The size of the Dutch subset reflected the larger coverage of the VanDale monolingual when compared to LDOCE, since the *drank* subset had 214 entries (compared to 192) and the *voedsel* subset 476 (compared to 260). Rather than comparing LKB entries, Vossen constructed tlinks from a database of dictionary entries augmented with the derived taxonomic information. The disambiguation depended entirely on the assumption that only one sense would be found within the target subset. The tlinks were generated completely automatically.

Vossen distinguished between four classes of tlink for this experiment:

**simple-tlink** This was applicable if a lexical entry from the source subset could be linked by the bilingual dictionary to a lexical entry found in the target subset. In the majority of cases, lexical entries will correspond to single orthographic words, although some multiword compounds are found, especially in LDOCE.

**orth-tlink** This was used for the cases where no entries were found in the bilingual dictionaries but the source and target subsets contained lexical entries with the same, or very similar orthography. These were nearly all cases where the entry corresponded to a loan word in one or both languages: e.g. *Chianti*. These **orth-tlink**s are essentially equivalent to **simple-tlinks**, but it was useful to record that they were derived without a bilingual dictionary.

**compound-tlink** There were many cases where translation fields contained Dutch compounds which were not present in the monolingual dictionary and hence not in the subset. For example, the English *creme de menthe* has the translation *pepermuntlikeur* in the VED, but *pepermuntlikeur* is not found as an entry in the VanDale monolingual. Dutch compounds can be morphologically complex, sometimes containing plural forms or a binding morpheme. These were analysed into parts which were checked for inclusion in the subset. Tlinks constructed in this way were given the type **compound-tlink**. Some examples are given below: note that these are not fully analysed since the part in brackets is not found in the subset and not disambiguated. Note also that this procedure generates multiple tlinks when more than one of the component parts is found in the subset, which may lead to incorrect results.

```
creme_de_menthe_L_0_0 / likeur_V_0_1 (pepermunt):
compound-tlink.

applejack_L_0_0 / wijn_V_0_1 (appelbrande):
compound-tlink.

applejack_L_0_0 / brandewijn_V_0_1 (appel):
```

*Table 1.* Dutch/English tlinks

| source subset | senses | simple tlinks | phrasal tlinks | compound tlinks | orth tlinks | % senses linked |
|---|---|---|---|---|---|---|
| Van Dale drank | 214 | 116 | 104 | 0 | 11 | 63% |
| Van Dale voedsel | 476 | 170 | 268 | 10 | 4 | 53% |
| LDOCE drink | 192 | 134 | 43 | 39 | 3 | 65% |
| LDOCE food | 260 | 72 | 124 | 114 | 14 | 55% |

`compound-tlink.`

**phrasal-tlink** These were treated similarly to compounds, resulting in tlinks such as

```
sarsaparilla_L_0_0 / frisdrank_V_0_1 (met sarsaparillasmaak):
phrasal-tlink.
```

Further examples of tlinks are given in the appendix. Both compound and phrasal tlinks as described here are essentially partial — their utilisation is discussed in §4, below. No rules have yet been implemented to deal with plural forms and derivations in translation fields.

The results for the four subsets are summarised in Table 1. In many cases multiple tlinks were produced for a sense, because multiple translations were given in the bilingual dictionary. Note that the senses not linked in this experiment could have been linked via their taxonomic parents, giving another form of partial tlink. This idea is described in more detail in the next section since it was extensively used in the Spanish experiment. The noticeably poorer linking rate in the food subsets is discussed below, in §4.

A small portion of the LKB entries could not be linked because of gaps in the bilingual dictionaries: going from English to Dutch 55 out of 452 senses, going from Dutch to English 66 out of 1050 senses. Most of these are true gaps in the bilingual dictionaries, e.g. *Bordeaux*, *green tea*, *Irish stew*, *moussaka*, *schuimbier*, *stortebier*, *dressing*. Some of these gaps are nevertheless resolved because of the orthographic equivalence heuristic. Some LKB entries have not been found because they were in an inflected form, e.g. *refreshments*, others have not been found because their entries are simple cross references e.g. *levensbenodigdheden* refers to *levensbehoefte* (essentials of life), *tarwebrood* refers to *tarwe* (wheat bread), but we would expect to be able to deal with most of these cases with a relatively small amount of work.

The expectation that polysemy in the target lexicon would be avoided by restricting it to a semantic subset was not entirely correct: there were 57 words with more than one sense in the combined subsets (out of a total number of 982). But, as discussed in [50], many of the sense distinctions involved are somewhat subtle. The sort of polysemy usually involved is exemplified by the LDOCE entry for *beer* where three senses are given: the first is the alcoholic drink sense, the second the corresponding conventional portion sense, and the third a *broadened* sense [15], which covers non-alcoholic *ginger beer*, etc. Although it is not particularly satisfactory to either rely entirely on manual intervention or to link all Dutch words which can be translated as *beer* to all three senses, the best approach to this problem would be to provide a more satisfactory monolingual representation of sense distinctions.

### 3.2. Spanish/English

The Spanish experiment was significantly different from the Dutch in a number of respects, perhaps most importantly in the nature of the bilingual dictionary used. The Vox-Harrap's Esencial [6], the only one available, is relatively small, containing about 16000 entries. By contrast, the monolingual VOX dictionary [5] used to generate the source subset contains almost 90000 entries. This results in omissions from the bilingual dictionary being a far more serious problem than in the preceding experiment.

Another major difference was that, in order to allow for convenient and flexible expansion of the heuristics to produce different types of tlinks in different situations, an interactive environment was developed in which to experiment with tlink construction. The Tlink Generation Environment (TGE) was implemented using a production rule approach which had already been utilised in the SEISD system (Sistema de Extracción de Información Semántica de Diccionarios: [1]) for constructing monolingual LKB entries from dictionaries. The core of the TGE is the production rule environment PRE, a rule-oriented general purpose interpreter adapted to natural language applications, which offers a powerful rule application mechanism allowing inheritance between rulesets and a choice of control strategies. TGE is described in more detail in [2], [3] and [38]. In contrast with the work on Dutch, tlinks can be checked by the user as they are being constructed.

The characterisation of tlink types used was slightly different from the previous experiment. Three main types were assumed: **simple-tlinks**, **phrasal-tlinks** and **partial-tlinks**, but **simple-tlinks** and **partial-tlinks** can be subdivided according to the TGE module or ruleset used to generate them.

**simple-tlink** As in the previous experiment, these correspond to mappings where there is a direct mapping between single entries in the source and target subsets.

- **simple tlink module** The mapping is generated from a word to word equivalence in the bilingual dictionary.
- **orthographic tlink module** This corresponds to the **orth-tlink** type of the previous experiment.
- **compound tlink module** This was applied where the English translation was made up of two or more orthographic words, but nevertheless corresponded to a single LKB entry (e.g. milk_shake_L_0_0).

**phrasal-tlink** These were distinguished from partial-tlinks and generated only when the phrase could be completely analysed. Appropriate rules have been developed for English noun-noun compounds and for verb-particle constructions. However there were very few cases where full analysis was possible for the drinks subset considered here.

**partial-tlink** Partial-tlinks were generated when the translation phrase in the bilingual dictionary could not be fully analysed, or where there are gaps in the bilingual dictionary or the target subset. (The phrasal and compound tlinks mentioned in §3.1 should be considered as varieties of partial tlink.) Three modules were developed to handle particular subcases:

**parent tlink module** Many terms in the monolingual source subset were not found in the bilingual dictionary, but their hyperonyms in the taxonomy had a clear translation. For example, although *coña* was not given in the bilingual dictionary, it could be linked to *cognac* in the English subset via a parent-link, because it was described in the Spanish monolingual dictionary using the genus term *coñac* which is translated as *cognac*.

**grandparent tlink module** This is similar to the previous case, but applies when the direct parent was also not translatable via the bilingual dictionary. (This happened quite frequently: the VOX monolingual is not a learners' dictionary and does not have a controlled definition vocabulary, so relatively uncommon words are sometimes found in definitions and the taxonomies tend to have more levels than those found in LDOCE.)

**general tlink module** This applies when a phrase which cannot be fully analysed is given as a translation in the bilingual dictionary. Like monolingual definitions, these can normally be regarded as consisting of a genus term and its modifiers, and the source entry can be linked to the entry for the genus term. The **compound-tlink** and **phrasal-tlink** types in the Dutch experiment essentially correspond to this case.

When using TGE for tlink generation, the lexicon builder must accept or reject the system's proposals. This process can be carried on in various ways depending on the TGE control strategy, the modules triggered by this strategy, and the overall goals of the procedure. TGE allows different modes of performance according to the intended goal. Each mode is appropriate for different kinds of coverage and implies different level of human intervention. It is possible, for instance, to apply all the generation modules in order to select all the possible tlinks, allowing the user to select the appropriate one(s). Another possibility is to rank the different rulesets and apply them in order, presenting potential tlinks one at a time to the user until one is selected. In the former case the user must select zero or more entries from a list, in the latter the choice is simply a yes/no question. There are also intermediate possibilities. Different tlink proposals can be ranked according to the LUCIFER appropriateness measure. However in the reported experiment the utility of this indicator was low, because of the lack of training material caused by the restricted amount of polysemy in the semantically restricted subsets.

*Table 2.* Spanish/English tlinks

| type | module | number of tlinks | Spanish entries | English entries |
|---|---|---|---|---|
| simple-tlink | simple | 41 | 26 | 31 |
| | compound | 1 | 1 | 1 |
| | orthographic | 13 | 13 | 13 |
| | total | 55 (14.5%) | | |
| phrasal-tlink | phrasal | 2 (0.5%) | 1 | 3 |
| partial-tlink | parent | 268 | 149 | 15 |
| | grandparent | 44 | 30 | 10 |
| | general | 8 | 7 | 6 |
| | total | 320 (85%) | | |

The task the user has to perform is thus quite simple but requires a good knowledge of both the source and target languages, although the system could provide all the information needed. Sometimes the decision to be taken requires that the user look at complementary information such as the source or target dictionary definitions, the source or target lexical entries, the bilingual information and so on. The need for this complementary information depends not only on the level of fluency of the user but also on the difficulty of the domain and the level of detail of the lexicons involved. In the experiment, whose results are given below, a rather conservative approach was taken, i.e. all the modules looking for tlinks were activated and all the generated tlinks acceptable in terms of translation and leading to differentiable target lexical entries were accepted as tlinks.

The Spanish drink taxonomy contained 235 senses and had 5 levels. In total 223 (95%) of these senses were linked, often by multiple tlinks. In fact, the total number of tlinks resulting from the experiment was considerably higher than the number of senses (377 tlinks for the 223 linked lexical entries), although most of these were partial links. The time taken by an expert lexicon builder to construct these tlinks was about three hours, i.e. about 2.1 tlinks/minute or 1.2 senses/minute. The results classified by tlink type are summarised in Table 2.

Only 74 source entries (31.5%) were found in the bilingual dictionary and another 13 were linked by comparison of orthography. This result is explained by the great number of highly specific lexical entries in the source subset, as is implied by the difference of coverage between the two dictionaries and the relatively high depth of the source taxonomy. The remaining 148 (63%) lexical entries could only be linked partially with this bilingual dictionary, even in the best case of no gaps in the monolingual English lexicon and complete analysis of phrases used in translation. In fact, the situation was worse than this, since out of 74 translations, only 30 directly corresponded to target LKB entries.

This experiment was carried out unidirectionally (unlike the Dutch trial) using Spanish as a source and English as a target. However it resulted in 52 English senses out of the 192 in the target subset being linked (27%). Of course these results must be viewed with caution. Some of the English senses linked by TGE correspond to partial tlinks and can, therefore, incorporate other information. In the next section some examples of this are discussed.

## 4. Discussion and evaluation

These trials demonstrate that senses in existing monolingual lexical knowledge bases of the sort developed in ACQUILEX can be linked automatically or with little human intervention, with the aid of bilingual MRDs. Expressed in percentages of senses linked, the results vary from 60% (English food subset) to 95% (Spanish drink subset). The quality of the links, that is, the percentage of simple rather than partial tlinks, varies considerably depending on the dictionaries. We would expect to be able to increase the rate of linking in the Dutch/English experiment to around the 95% figure by making use of parent tlinks as in the Spanish/English trials. Of course the technique would be useful with much lower linking rates, because it automates work that would otherwise have to be done manually.

From our perspective these results are a success, since they demonstrate that much of the work involved in creating multilingual LKBs from monolingual LKBs could be automated, despite the sense correspondence problem. As a technique for building lexicons for MT, there are several qualifications. Firstly, this work requires existing semantically organised monolingual LKBs. Such LKBs are independently desirable, but although they can be constructed semi-automatically from MRDs, the amount of human intervention involved is significant (although less so for LDOCE than the other dictionaries). Secondly, although as indicated in §2 some complex cases of translation equivalence can be derived from simple tlinks, there are many complex mappings which cannot be constructed in this way and which cannot feasibly be constructed automatically from MRDs. The trials reported here all involve concrete nouns, and we would expect abstract nouns, verbs and adjectives to be considerably more difficult. These results therefore show potential for extending the vocabulary of an MT system, but we would not expect to automatically construct the core bilingual lexicon. Thirdly, the high linking rates involve the use of partial tlinks of various types. We believe these can be utilised in MT, but we return to this point below, after discussing some other issues raised by these experiments.

One limitation of the work described above is that the food and drink taxonomies do not contain all the senses which should be included in the respective subsets, because lexicographers chose to classify them under another genus term.[9] For example, no types of cheese were included in the Dutch *voedsel* subset because *kaas* comes under *zuivelproduckt* (dairy product) in the taxonomy and not under *voedsel*. Table 3 compares our best estimate for the size of the lexicons containing all senses corresponding to possible foods and drinks for LDOCE and VanDale with the size

*Table 3.* Failure rate by subset for Dutch/English tlinks

| | taxonomy size | full subset size | % tlinks missed |
|---|---|---|---|
| LDOCE food | 260 | 594 | 40 |
| LDOCE drink | 192 | 202 | 16 |
| Van Dale food | 476 | 1190 | 18 |
| Van Dale drink | 214 | 261 | 6 |

of the taxonomies, and gives the percentage of entries in each subset which were not linked by simple tlinks because of classification problems.

It can thus be seen that, although differences in classification were not the major source of gaps overall, they were responsible for the poorer linking rate in the food subsets in the Dutch experiments compared to the drink subset. The LDOCE classifications are more systematic than those from Van Dale in that a smaller percentage of senses are found outside the taxonomies. This probably reflects the restricting effect the use of a core defining vocabulary in LDOCE had on the taxonomies. But there were some examples of genuine conceptual mismatches between the two subsets, for example, Dutch *drank* also included liquid medicines and so on, such as *mondspoeling* (*mouthwash*) and *purgeerdrank* (*laxative, purgative*). It is thus clear that results would improve if we were in a position to run experiments over the entire dictionaries. The type hierarchy in the LKB does not reflect the taxonomies directly: for example *zuivelproduckt* would be classified as having the type which covers foods and drinks, and so this set of omissions would have been avoided if the subset had been defined in terms of LKB types, rather than taxonomically. Because of this, and because of the polysemy problems mentioned previously, the strategy adopted in the Spanish experiment of comparing LKB entries rather than derived dictionary entries is preferable in the longer term.

We mentioned in §2 that the tlink formalism was designed to be symmetric, but it is clear from examination of the results that different tlinks are produced by going from Dutch to English via the VDE, than from English to Dutch using the VED. Clearly lexicographers are not going to give infrequently used words or senses as translations if alternatives are available, though they have to list the obscure words if the dictionary is intended to be comprehensive. For example, both *cider* and *cyder* are found in the VED but we find *cider* but not *cyder* given as a translation of *appelwijn* in the VDE. This effect partly accounts for the low percentage of linking of English senses in the Spanish experiment which only used a Spanish-English bilingual dictionary: the less common English words were not used as translations and were therefore not linked. For practical purposes, it would obviously be better to prefer tlinks which have the 'right' directionality in translation, but since this is a preference, not an absolute restriction, it does not affect our basic assumption of bidirectionality. It would also be sensible to prefer the main translations given in the VDE and the VED over the secondary translations. In general, however, we regard lexical choice as an issue which is distinct from acquisition and construction of the lexicons, since we see it as important for domain independence and maximum flexibility to give a range of translation possibilities.

It turns out that some senses which we should classify as translational equivalents do not have compatible RQSs. One example, discussed in detail by Vossen [50] is *Sauternes*, which has the following definitions in LDOCE and Van Dale respectively.

**LDOCE** a type of sweet gold-coloured French wine

**VanDale** Franse zoete, witte wijn, vooral bij vis gebruikt
French sweet, white wine, esp. used with fish

The English RQS has the feature COLOUR instantiated to **gold**, whereas for Dutch it is **white**. Now, intuitively, *Sauternes* is a clear case of a term that should be translation equivalent between English and Dutch. It is a loan word in both languages, and within the EU it has an imposed definition: it refers to a certain type of wine from a particular region and it would be illegal to use the term to label any other drink. There thus seems more possibility of a misunderstanding between speakers of British and American English than between Dutch and British English speakers (since in the USA all sorts of sweet white wine are described as *Sauternes*). The discrepancy in the RQS could be described as a representational failure: it would not be contradictory to refer to *Sauternes* as *a gold-coloured white wine* since *white* is a classification of wine which has a rather indirect relationship to perceived colour (and the same is true of Dutch *wit*). So the COLOUR feature of the Dutch RQS should perhaps not have been instantiated to **white**. This illustrates the difficulty of constructing an adequate semantic representation: since we cannot expect to get this completely correct, we have to accept that some minor discrepancies will occur.[10]

In general, however, we would expect RQSs of translation-equivalent structures to be unifiable, though not identical since different lexicographers will choose to highlight different aspects of a concept, and, in any case, our automatic analysis of the definitions is sometimes incomplete. We can thus transfer RQS information between the monolingual concepts. Although we could avoid errors due to mismatch by using default rather than monotonic unification, it is probably better to signal the discrepancies and allow the user to resolve them manually, as appropriate.

We have not had the opportunity to investigate the use of derived tlinks in a full MT system. Trujillo [46] has developed an approach to MT using a bilingual lexicon as the only transfer component: the formalism assumed is similar to tlinks and the system is implemented using the ACQUILEX LKB framework. Trujillo demonstrates the adequacy of his representation for translation of locative prepositions, making use of a semantic classification of nouns which is compatible with the assumptions we have made about the RQS. There are two main issues to consider with respect to the utilisation of tlinks in more established approaches to MT: whether the information encoded in the tlinks (mainly via the monolingual lexical entries) is rich enough to support various MT approaches, and whether the formalism allows the tlinks to be appropriately translated. Full tlinks contain more information than is assumed by most approaches to transfer based translation, as illustrated by the bilingual lexical entries given as in papers such as those by Estival *et al.* [19] and Alshawi *et al.* [4]. Formally, there is a close correspondance between the equivalence of variables in tlinks and the use of transfer variables assumed in those approaches. Furthermore, because the tlink description abstracts away from details of the linguistic encoding as much as possible, by expressing equivalences in terms of the monolingual lexicon, it would be relatively easy to automatically transform at least the simpler tlinks into the bilingual lexicon needed. The main issue would be the compatability of the syntactic representation, but this is not particularly problematic for nouns. Most of the semantic detail would be simply discarded. It is more difficult to see whether tlinks have anything to offer knowledge based approaches to MT, because, although the amount of detail in the noun representations appears at least comparable to those illustrated in Nirenburg *et al.*[33], for example, there is far less agreement about semantic ontologies than syntactic categories.

However we think the main issue, at least for nouns, is how a translation system can make use of partial tlinks. As we saw above, there were many cases where the only tlink that could be constructed was a partial-tlink of some type. In a few cases, this indicates a problem with the methodology: for example, Spanish *zumo* could only be given a partial tlink to *drink* rather than *juice*, because *juice* was missing from the drink taxonomy in English. But most cases are indicative of gaps, either in coverage of the MRDs or of a real lack of a direct translation into the target language. In some cases the concepts involved are 'local', in that they denote a regional product or speciality, for example. For example, *Penedés*, *Rioja* and *Malaga* are in the Spanish drink subset as types of wine from those localities. Many more types of gin are available in the Netherlands than in Britain, so we find *dubbelgebeide* (*gin made with twice the normal quantity of juniper berries: prime quality gin*), *bessejenever* (*blackcurrant gin*), *moutjenever* (*malt gin*) and so on.

Whether we find such words in a bilingual dictionary or not will depend on its size, but the translation given is normally a phrase, which explains or paraphrases the concept and is therefore not necessarily appropriate as a translation in a given context. For example, (3a) would be distinctly odd as part of a narrative and either (3b) or (3c) would probably be better.

(3) a He went into the bar and ordered a gin made with twice the normal quantity of juniper berries.
b He went into the bar and ordered a gin.
c He went into the bar and ordered a *dubbelgebeide*.

Thus even good, wide-coverage bilingual dictionaries, such as the Van Dale ones, do not always give translations that are usable. The choice of food and drink as subsets means that the genuine mismatches we find are normally because of cultural localisation, rather than differences of conceptualisation between two languages. But the same effect is observed with many other classes of noun (professions, institutions, transport, etc.), and it is essential to be able to deal with such cases in a broad coverage MT system (and in machine-aided translation).

Our use of partial tlinks for such cases is not just an artifact of our extraction methodology but represents a first step in addressing this problem, which will arise for machine translation lexicons however they are generated. As we said in the introduction, the use of RQS can be seen as a way of partially decomposing the semantic space for a particular class of senses. Some such concepts correspond to a single word (or established compound, etc.) in one language but cannot be expressed (even approximately) without the use of a phrase in another. Rather than hard-wiring a particular phrase into the bilingual lexicon, the representation gives us the option of constructing a partial tlink, and transferring the extra information from the source monolingual entry to the target representation, giving a feature structure which incorporates the additional semantic information which the target generator can choose to express or not, as appropriate. For example, the Spanish *Penedés* can be linked to a FS equivalent to that for *wine* minus the orthography with the ORIGIN-AREA feature in the RQS instantiated to **Penedés**. In this particular case, there is a lexical rule for English that a place name can also be used to refer to a characteristic product of that place. Thus if *Penedés* is accessible to the English generator as a place name the lexical rule will generate an underspecified representation which will unify with the output of the tlink (ignoring orthography).[11] Thus, the generator will find the string "Penedés" as a possible realisation of the semantics, even though it is not an explicit entry in the target lexicon. Alternatively the modifier might be ignored altogether or a phrase *wine from Penedés* might be constructed. (Since the RQSs are derived by parsing the monolingual descriptions, we could in principle reverse this process, allowing the generator to produce definition-like strings as one mode of operation.)

Of course, this requires that the target generator is capable of recognising whether an underspecified translation is appropriate or not. In the limit, this is far beyond the capabilities of today's systems, since it would require detailed knowledge of the author's intentions. But there are contexts where it is possible to think of automatically applicable rules. A bilingual restaurant menu, for example, should have a paraphrase rather than a generic term or repetition of the source language word, because it is appropriate for it to be maximally informative. In contrast, a translation into English of an article about Spanish wine, for example, should refer to *Penedés* rather than *wine from Penedés* to avoid redundancy. In general, we see partial tlinks as providing the information necessary to make such choices possible as an alternative to hardwiring a particular translation. This fits in naturally with the approach of translation by negotiation advocated by Kay [23]: the generator can either accept the partial information as adequate in the context, or attempt to construct a target language representation which conveys as much of the extra information as is needed. From the lexical viewpoint, the problem comes in attempting to encode the right information in the partial interlingua, currently embodied by the RQS, to make such flexibility possible, without sacrificing efficiency and directness in the (many) cases where a straightforward lexical transfer is appropriate.

## 5. Conclusion

The experiments reported have shown the feasibility of constructing useful fragments of multilingual lexicons from MRDs but have also demonstrated the limitations of existing dictionaries. Although there are various improvements that we could make to the extraction technology in order to avoid construction of spurious tlinks, there is no prospect of constructing completely adequate lexicons fully automatically from MRDs, because the dictionaries themselves do not contain all the necessary information. As in the monolingual case, MRDs are most usefully seen as tools to speed production of the lexicon, or to extend a core lexicon, not to automate lexicon construction entirely. Similarly, linked LKBs have potential for machine-aided translation, to give a human translator a possible translation or choice of translations efficiently, omitting terms which would appear in a bilingual dictionary but which are syntactically or semantically inappropriate in that context.

There are several areas of further work, which we have started or intend to carry out. In order to improve the sense-to-sense mapping, a more satisfactory representation of polysemy in the monolingual lexicons is needed than treating each sense given by the lexicographer as fully distinct: some theoretical work on polysemy is described in [15] and we expect this to lead to an improved representation within the LKB. Systematic translation mismatches are being investigated for Spanish and English [27].

Further experiments on tlink generation will be carried out on other syntactic and semantic categories, but this requires the use of larger monolingual lexicons. Extension of the current type system to include adjectives and a finer-grained semantic classification for verbs (verbal RQS) is currently being carried out within the ACQUILEX II project. The definition of selectional restrictions in order to generate phrasal tlinks and the extension of tlinks from lexical entries to lexical rules is currently under research. Special attention has been paid to the relationships between verbal diathesis rules in Spanish-English.

One final point is that, to some extent, the inadequacies of bilingual dictionaries as sources for MT lexicons, also reflect inadequacies in their intended use for humans. For example, the Deutsch-Englisch Langenscheidts Taschenwörterbücher [30], contains the entry:

**Kieme** *f* gill.

with no indication of which sense of *gill* is meant. To a British English speaker, at least, this could plausibly refer to a slit which is part of a fish or a mushroom, a measure of volume for liquids, or a fast running stream. The sort of techniques which we have described here for constructing tlinks could be adapted to automatically indicate the possibility of ambiguity to a lexicographer who creates such an entry, prompting them to add a disambiguating word or label. There are obvious advantages in both monolingual and bilingual cases for incorporating a semantic coding scheme to be instantiated by lexicographers, since this would allow consistency and coherence checks of this sort to be carried out, and allow systematic sense extensions and translation mismatches to be dealt with consistently. Such techniques are actively under investigation by dictionary publishers in the context of the ACQUILEX II project with the aim not only of improving printed dictionaries, but of creating lexical databases which can also be exploited for NLP and MT. The long-term utility of the work reported here almost certainly rests on utilising the insights gained to help determine the content and format of these lexical databases so that many of the problems identified are resolved when they are constructed.

### Acknowledgments

This work was carried out under the ACQUILEX projects (3030 and 7315) funded by the EC under the Esprit basic research program. Francesc Ribas was supported by a grant from the Departament d'Ensenyament of Generalitat de Catalunya, 91-DOCG-1491. German Rigau was supported by a grant from the Ministerio de Educación y Ciencia, 92-BOE-16392. We are grateful to the publishers, Van Dale, Biblograf and Longman for giving us permission to use their dictionaries and to Arturo Trujillo and two anonymous referees for their comments on this paper.

### Appendix

For reasons of space the tlinks shown are limited to those with source entries beginning with a or b, except where this would have excluded all members of a class, in which case the first one or two entries of the class are shown.

### Dutch drank to English drink

**Simple tlinks**
ABSINT_V_0_1/ ABSINTH_L_0_1
ADVOCAAT_V_0_3/ EGGNOG_L_0_0
ADVOCAAT_V_0_3/ FLIP_L_2_2
ALCOHOL_V_0_2/ ALCOHOL_L_0_2
ALCOHOL_V_0_2/ DRINK_L_2_1
ALCOHOL_V_0_2/ SPIRIT_L_1_12
APPELWIJN_V_0_1/ CIDER_L_0_1
APPELWIJN_V_0_1/ CIDER_L_0_3
ARAK_V_0_1/ ARRACK_L_0_0
BEAUJOLAIS_V_0_1/ BEAUJOLAIS_L_0_0
BIER_V_0_1/ ALE_L_0_0
BIER_V_0_1/ BEER_L_0_1
BIER_V_0_1/ BEER_L_0_2
BIER_V_0_1/ BEER_L_0_3

BISSCHOPWIJN_V_0_1/ WINE_L_1_1
BISSCHOPWIJN_V_0_1/ WINE_L_1_2
BOURBON_V_0_1/ BOURBON_L_0_0
BOURGOGNE_V_0_1/ BURGUNDY_L_0_0
BOWL_V_0_2/ CUP_L_1_6
BOWL_V_0_2/ PUNCH_L_4_0
BRANDEWIJN_V_0_1/ BRANDY_L_1_1
BRANDEWIJN_V_0_1/ BRANDY_L_1_2

**Phrasal tlinks**

BESSENJENEVER_V_0_1/ GIN_L_2_0 (BLACKCURRANT)
BESSENWIJN_V_0_1/ WINE_L_1_2 (CURRANT)
BESSENWIJN_V_0_1/ WINE_L_1_1 (CURRANT)
BESSENWIJN_V_0_1/ WINE_L_1_2 (RED CURRANT)
BESSENWIJN_V_0_1/ WINE_L_1_1 (RED CURRANT)
BITTER_V_0_1/ GIN_L_2_0 (AND BITTERS)
BLOEDWIJN_V_0_1/ WINE_L_1_2 (TONIC)
BLOEDWIJN_V_0_1/ WINE_L_1_1 (TONIC)
BOLS_V_0_1/ GIN_L_2_0 (DUTCH)

**Orthographic tlinks**

BITTER_V_0_1/ BITTER_L_3_0
BRANDY_V_0_1/ BRANDY_L_1_2
BRANDY_V_0_1/ BRANDY_L_1_1

**Tlinks derived from English drink to Dutch drank**

**Simple Tlinks**

ABSINTH_L_0_1/ ABSINT_V_0_1
ALCOHOL_L_0_2/ ALCOHOL_V_0_2
ALE_L_0_0/ BIER_V_0_1
AQUA_VITAE_L_0_0/ ALCOHOL_V_0_2
AQUA_VITAE_L_0_0/ BRANDEWIJN_V_0_1
ARRACK_L_0_0/ ARAK_V_0_1
BEAUJOLAIS_L_0_0/ BEAUJOLAIS_V_0_1
BEER_L_0_1/ BIER_V_0_1
BEER_L_0_2/ BIER_V_0_1
BEER_L_0_3/ BIER_V_0_1
BITTER_L_3_0/ BITTER_V_0_1
BOCK_L_0_0/ BOCKBIER_V_0_1
BOOZE_L_2_1/ DRANK_V_0_1
BOOZE_L_2_1/ DRANK_V_0_2
BOOZE_L_2_1/ STERKEDRANK_V_0_1
BOTTLE_L_1_3/ DRANK_V_0_1
BOTTLE_L_1_3/ DRANK_V_0_2
BOURBON_L_0_0/ BOURBON_V_0_1

BRANDY_L_1_1/ BRANDEWIJN_V_0_1
BRANDY_L_1_1/ COGNAC_V_0_1
BRANDY_L_1_2/ BRANDEWIJN_V_0_1
BRANDY_L_1_2/ COGNAC_V_0_1
BURGUNDY_L_0_0/ BOURGOGNE_V_0_1

**Phrasal Tlinks**

BLACK_AND_TAN_L_0_3/ BITTER_V_0_1 (MENGSEL V BIER EN STOUT)
BLACK_AND_TAN_L_0_3/ BIER_V_0_1 (MENGSEL V BITTER EN STOUT)
BLACK_AND_TAN_L_0_3/ STOUT_V_0_1 (MENGSEL V BITTER BIER EN)

**Compound Tlinks**

ANIS_L_0_0/ LIKEUR_V_0_1 (ANIJS)
APPLEJACK_L_0_0/ WIJN_V_0_3 (APPELBRANDE)
APPLEJACK_L_0_0/ WIJN_V_0_1 (APPELBRANDE)
APPLEJACK_L_0_0/ BRANDEWIJN_V_0_1 (APPEL)
BARLEY_WINE_L_0_0/ WIJN_V_0_3 (GERSTE)
BARLEY_WINE_L_0_0/ WIJN_V_0_1 (GERSTE)

**Orthographic Tlinks**

COLA_L_0_0/ COLA_V_0_1

**Spanish bebida to English drink**

**Simple tlink module**

AGUARDIENTE_X_I_1 / LIQUOR_L_0_1
AGUARDIENTE_X_I_1 / LIQUOR_L_0_2
AGUARDIENTE_X_I_1 / BRANDY_L_1_1
AGUARDIENTE_X_I_1 / BRANDY_L_1_2
AJENJO_X_I_2 / ABSINTH_L_0_1
BEBEDIZO_X_I_2 / POTION_L_0_0
BEBEDIZO_X_I_3 / POTION_L_0_0
BEBIDA_X_I_3 / DRINK_L_2_1
BRANDI_X_I_1 / BRANDY_L_1_1
BRANDI_X_I_1 / BRANDY_L_1_2

**Compound tlink module**

BATIDO_X_I_5 / MILK_SHAKE_L_0_0

**Orthographic tlink module**

ALE_X_I_1 / ALE_L_0_0
BOURBON_X_I_1 / BOURBON_L_0_0

**Parent tlink module**

ABSENTA_X_I_1 / ABSINTH_L_0_1
ABSENTA_X_I_1 / DRINK_L_2_1
ADIAFA_X_I_1 / REFRESHMENT_L_0_2
AGASAJO_X_I_3 / REFRESHMENT_L_0_2
AGUACHIRLE_X_I_1 / WINE_L_1_1
AGUACHIRLE_X_I_1 / WINE_L_1_2

AGUAPIÉ_X_I_1 / WINE_L_1_1
AGUAPIÉ_X_I_1 / WINE_L_1_2
AHUMADO_X_I_4 / WINE_L_1_1
AHUMADO_X_I_4 / WINE_L_1_2
ALBARIÑO_X_I_1 / WINE_L_1_1
ALBARIÑO_X_I_1 / WINE_L_1_2
ALICANTE_X_I_3 / WINE_L_1_1
ALICANTE_X_I_3 / WINE_L_1_2
ALMENDRADA_X_I_1 / DRINK_L_2_1
ALOQUE_X_I_2 / WINE_L_1_1
ALOQUE_X_I_2 / WINE_L_1_2
ALQUERMES_X_I_2 / LIQUOR_L_0_1
ALQUERMES_X_I_2 / LIQUOR_L_0_2
AMONTILLADO_X_I_1 / WINE_L_1_1
AMONTILLADO_X_I_1 / WINE_L_1_2
AMOROSO_X_I_5 / WINE_L_1_1
AMOROSO_X_I_5 / WINE_L_1_2
ANGÉLICA_X_I_5 / DRINK_L_2_1
ANISADO_X_I_2 / BRANDY_L_1_1
ANISADO_X_I_2 / BRANDY_L_1_2
ANISADO_X_I_2 / LIQUOR_L_0_1
ANISADO_X_I_2 / LIQUOR_L_0_2
ANISETE_X_I_1 / LIQUOR_L_0_1
ANISETE_X_I_1 / LIQUOR_L_0_2
AÑEJO_X_I_3 / WINE_L_1_1
AÑEJO_X_I_3 / WINE_L_1_2
APERITIVO_X_I_2 / DRINK_L_2_1
ARLEQUÍN_X_I_5 / DRINK_L_2_1
ARMAÑAC_X_I_1 / BRANDY_L_1_1
ARMAÑAC_X_I_1 / BRANDY_L_1_2
ARMAÑAC_X_I_1 / LIQUOR_L_0_1
ARMAÑAC_X_I_1 / LIQUOR_L_0_2
ATOLE_X_I_1 / DRINK_L_2_1
AURORA_X_I_6 / DRINK_L_2_1
AVENATE_X_I_1 / DRINK_L_2_1
BALARRASA_X_I_1 / BRANDY_L_1_1
BALARRASA_X_I_1 / BRANDY_L_1_2
BALARRASA_X_I_1 / LIQUOR_L_0_1
BALARRASA_X_I_1 / LIQUOR_L_0_2
BÁSIG_X_I_1 / LIQUOR_L_0_1
BÁSIG_X_I_1 / LIQUOR_L_0_2
BEBER_X_I_1 / DRINK_L_2_1:
BEBIDO_X_I_2 / DRINK_L_2_1
BEBIENDA_X_I_1 / DRINK_L_2_1

BENEDICTINO_X_I_4 / LIQUOR_L_0_1
BENEDICTINO_X_I_4 / LIQUOR_L_0_2
BREBAJE_X_I_1 / DRINK_L_2_1
BÍTER_X_I_1 / DRINK_L_2_1
**Grandparent tlink module**
ACETOMIEL_X_I_1 / DRINK_L_2_1
ANÍS_X_I_5 / BRANDY_L_1_1
ANÍS_X_I_5 / BRANDY_L_1_2
ANÍS_X_I_5 / LIQUOR_L_0_1
ANÍS_X_I_5 / LIQUOR_L_0_2
ARLEQUÍN_X_I_5 / REFRESHMENT_L_0_2
ARROPE_X_I_3 / DRINK_L_2_1
BREBAJO_X_I_1 / DRINK_L_2_1
BÍTTER_X_I_1 / DRINK_L_2_1
**General tlink module**
ESTOMACAL_X_I_2 / LIQUEUR_L_0_0
FILTRO_X_II_2 / POTION_L_0_0
**Phrasal noun tlink module**
CLARETE_X_I_1 / CLARET_L_1_0+WINE_L_1_1
CLARETE_X_I_1 / CLARET_L_1_0+WINE_L_1_2

## Notes

1. Further details of the work reported in this paper can be found in [3] and in project working papers [50], [2], [38], [16].
2. We have, on the whole, perpetuated the tradition of using English words to name language independent concepts, partly because of the central status of the English monolingual lexicon in the project and partly because it is the language most project members have in common.
3. Sutcliffe [44] describes the use of weighted features bundles to represent information extracted from MRDs and suggests that comparison of these alone can be used to construct a bilingual lexicon, by linking entries with the most similar set of features.
4. Tlinks are both symmetrical and reversible; we use the terminology source, target, input and output solely for ease of exposition.
5. The morphological generalisation involved in this example can be captured, making it unnecessary to manually specify both tlinks.
6. In some cases diminutive formation in Dutch appears to have the effect of portioning, for example, *biertje* can be translated as *glass of beer* [48].
7. However, because it appears that there are a limited number of possible paths that can be expressed in natural language, it would perhaps be more elegant to make use of a partial interlingua here, and relate both *across* and *cruzar* to a single underlying primitive expressing the path in the monolingual lexicon. This would enable us to avoid having to explicitly specify the *across/cruzar* link.
8. Complex translation fields, where information is conflated according to various lexicographic conventions, are excluded from this figure but were parsed for the tlink construction experiment.
9. Vossen and Copestake [51] discuss the classification found in LDOCE which results in *vegetable*, *cream*, *meat* and *spice* all being found outside the food taxonomy.
10. The first named author is more disturbed by *vooral bij vis gebruikt*: *esp. used with fish*, which, if it is not an aberration on the part of an individual lexicographer, indicates a significant difference between Dutch and English tastes.
11. Note that the output of the lexical rule is underspecified, because we do not assume that the lexicon contains information about what is produced in Penedés.

## References


1. Ageno, A., I. Castellon, G. Rigau, H. Rodriguez, M.F. Verdejo, M.A. Marti and M. Taule (1992) 'SEISD: An Environment for Extraction of Semantic Information from On-Line Dictionaries', *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92),* Trento, Italy, pp. 253–255.
2. Ageno, A., F. Ribas, G. Rigau, H. Rodriguez, F.Verdejo (1993) *TGE: Tlinks Generation Environment,* ACQUILEX II Working paper 8.
3. Ageno, A., F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou (1994) 'TGE: Tlinks Generation Environment', *Proceedings of the 15th International Congress on Computational Linguistics (COLING 94),* Kyoto, Japan.
4. 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.
5. Biblograf S.A.(publisher) (Biblograf, S.A.: Barcelona) *Diccionario general ilustrado de la Lengua Española VOX,* 1987
6. Biblograf S.A. (publisher) (Biblograf, S.A.: Barcelona) *VOX Harraps Diccionario Esencial: Ingés-Español Español-Inglés. Second edition.,* 1992
7. Boguraev, B. and E. J. Briscoe (eds) (1989) *Computational lexicography for natural language processing,* Longman, London.
8. Briscoe, E. J. (1991) 'Lexical issues in natural language processing' in E. Klein and F. Veltman (eds.), *Natural language and speech,* Springer-Verlag, pp. 39–68.

9. Calzolari, N. (1992) 'Acquiring and representing semantic information in a lexical knowledge base' in J. Pustejovsky and S. Bergler (eds.), *Lexical Semantics and Knowledge Representation,* Lecture Notes in AI: 627, Springer-Verlag, Berlin, pp. 235–244.
10. Carpenter, R. (1992) *The Logic of Typed Feature Structures,* Cambridge University Press, Tracts in Theoretical Computer Science.
11. Carroll, J. and C. Grover (1989) 'The derivation of a large computational lexicon for English from LDOCE' in B. Boguraev and E. J. Briscoe (eds.), *Computational lexicography for natural language processing,* Longman, London, pp. 117–134.
12. Copestake, A. (1990) 'An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary', *Proceedings of the First International Workshop on Inheritance in Natural Language Processing,* Tilburg, The Netherlands, pp. 19–29.
13. 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.
14. Copestake, A. (1993) *The Compleat LKB,* University of Cambridge Computer Laboratory Technical Report No. 316.
15. Copestake, A. and E.J. Briscoe (1995) 'Semi-Productive Polysemy and Sense Extension', *Journal of Semantics, vol.12,* 15–67.
16. 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.
17. Copestake, A. and A. Sanfilippo (1993) 'Multilingual lexical representation', *Proceedings of the AAAI Spring Symposium on Building lexicons for machine translation,* Stanford, CA.
18. Dowty, D. (1991) 'Thematic proto-roles and argument selection ', *Language, vol.67,* 547–619.
19. 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.
20. Helmreich, S., L. Guthrie and Y. Wilks (1993) *The Use of Machine Readable Dictionaries in the Pangloss Project,* Paper presented at the AAAI Spring Symposium on Building Lexicons for Machine Translation, Stanford University.
21. Hutchins, W.J. and H.L. Somers (1992) *An Introduction to Machine Translation,* Academic Press, London.
22. 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.
23. Kay, M., J-M. Gawron and P. Norvig (1994) *Verbmobil: A translation system for face-to-face dialog,* CSLI Lecture Notes, No. 33: University of Chicago Press.
24. Klavans, J.L. (1988) 'COMPLEX: A Computational Lexicon for Natural Language Systems', *Proceedings of the 12th International Conference on Computational Linguistics (Coling-88),* Budapest, Hungary.
25. Klavans, J.L., M.S. Chodorow and N. Wacholder (1990) 'From Dictionary to Knowledge Base via Taxonomy', *Proceedings of the Sixth Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research: Electronic Text Research,* University of Waterloo, Canada.
26. Levin, B. (1993) *English Verb Classes and Alternations,* Chicago University Press, Chicago.
27. Marti, M.A. and C. Soler (1994) 'Dealing with lexical mismatches', *Proceedings of the Euralex 94,* Amsterdam, The Netherlands.
28. Martin, W. and G.A.J. Tops (1984) *Groot Woordenboek Engels-Nederlands,* Van Dale Lexicografie: Utrecht, The Netherlands.
29. Martin, W. and G.A.J. Tops (1986) *Groot Woordenboek Nederlands-Engels,* Van Dale Lexicografie: Utrecht, The Netherlands.
30. Messinger, H. and G. Türck (1990) *Deutsch-Englisch Langenscheidts Taschenwörterbücher,* Langenscheidt, Berlin.

31. Neff, M., B. Blaser, J-M Lange, H. Lehmann and I. Dominguez (1993) *Get it where you can: Acquiring and Maintaining Bilingual Lexicons for Machine Translation,* Paper presented at the AAAI Spring Symposium on Building Lexicons for Machine Translation, Stanford University.
32. Neff, M. and M. McCord (1990) 'Acquiring Lexical Data from Machine-Readable Dictionary Resources for Machine Translation', *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-90),* Austin, Texas.
33. Nirenburg, S., V. Raskin and A. B. Tucker (1987) 'The Structure of Interlingua in TRANSLATOR' in S. Nirenburg (eds.), *Machine Translation: Theoretical and Methodological Issues,* Cambridge University Press, Cambridge, England, pp. 90–113.
34. Nirenburg, S. and V. Raskin (1988) 'The Subworld Concept Lexicon and the Lexicon Management System', *Computational Linguistics, vol.13(3-4),* 276–289.
35. Procter, P. (ed) (1978) *Longman Dictionary of Contemporary English,* Longman, London.
36. Pustejovsky, J. (1991) 'The generative lexicon', *Computational Linguistics, vol.17(4),* 409–441.
37. Samiotou, A. (1993) *Performance of Cross-linguistic Equivalence Relations: A Lexicon-based approach,* MSc Dissertation, Centre for Computational Linguistics, UMIST, England.
38. Samiotou, A., I. Castellon, F. Ribas, G. Rigau (1994) *Translation Equivalence via Lexicon: A Study on Tlinks,* ACQUILEX II Working paper 25.
39. Sanfilippo, A. (1993) 'LKB encoding of lexical knowledge from machine-readable dictionaries' in E. J. Briscoe, A. Copestake and V. de Paiva (eds.), *Inheritance, defaults and the lexicon,* Cambridge University Press, Cambridge, England, pp. 190–222.
40. 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.
41. 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.
42. Shieber, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar,* CSLI Lecture Notes 4, Stanford CA.
43. van Sterkenburg P. and W.J.J. Pijnenburg (1984) *Groot woordenboek van hedendaags Nederlands,* Van Dale Lexicografie: Utrecht, The Netherlands.
44. Sutcliffe, R. (1993) *Using Distributed Patterns as Language Independent Lexical Representations,* Paper presented at the AAAI Spring Symposium on Building Lexicons for Machine Translation, Stanford University.
45. Talmy, L. (1985) 'Lexicalization patterns' in T. Shopen (eds.), *Language Typology and Syntactic Description,* Cambridge University Press.
46. Trujillo, I.A. (1995) *Machine Translation with the ACQUILEX LKB,* ACQUILEX II Working Paper.
47. Vossen, P. (1990) *A Parser-Grammar for the Meaning Descriptions of LDOCE,* Links Project Technical Report 300-169-007, Amsterdam University.
48. Vossen, P. (1991) *Converting Data from a Lexical Database to a Knowledge Base,* Acquilex working paper 27: University of Amsterdam.
49. Vossen, P. (1992) 'The automatic construction of a knowledge base from dictionaries: a combination of techniques' in H. Tommola, K. Tarantola, T. Salmin Tolonen, J. Schop (eds.), *Euralex '92 Proceedings I-II,* Tampere, Finland.
50. Vossen, P. (1993) *Extracting equivalence relations for a multilingual Lexical knowledge base,* Acquilex 2 working paper No. 14: University of Amsterdam.
51. Vossen, P. and A. Copestake (1993) 'Untangling definition structure into knowledge representation' in E. J. Briscoe, A. Copestake and V. de Paiva (eds.), *Defaults, Inheritance and the Lexicon,* Cambridge University Press, Cambridge, England, pp. 246–274.
52. Vossen, P., W. Meijs and M. Broeder (1989) 'Meaning and structure in dictionary definitions' in B. Boguraev and E.J. Briscoe (eds.), *Computational Lexicography for Natural Language Processing,* Longman, London, pp. 171–192.

53. Walker, D. and R. Amsler (1986) 'The use of machine-readable dictionaries in sublanguage analysis' in R. Grishman and R. Kittredge (eds.), *Analyzing Language in Restricted Domains,* Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp. 69–83.
54. Whitelock, P. (1992) 'Shake-and-bake translation', *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92),* Nantes, France.
55. Wilks, Y., D. Fass, C-M. Guo, J. McDonald, T. Plate and B. Slator (1989) 'A tractable machine dictionary as a resource for computational semantics' in B. Boguraev and E. J. Briscoe (eds.), *Computational lexicography for natural language processing,* Longman, London, pp. 193–231.
56. 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.
57. Zeevat, H., E. Klein and J. Calder (1987) 'An introduction to unification categorial grammar' in N. Haddock, E. Klein and G. Morrill (eds.), *Categorial grammar, unification grammar, and parsing: working papers in cognitive science, Volume 1,* Centre for Cognitive Science, University of Edinburgh, pp. 195–222.

(ACQUILEX II Working Papers are available from cide@cup.cam.ac.uk)