TGE: Tlinks Generation Environment.

Alicia Ageno, Francesc Ribas[1], German Rigau[2], Horacio Rodríguez, Anna Samiotou.

Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. horacio@lsi.upc.es

Abstract

This paper describes the enhancements made, within a unification framework, based on typed feature structures, in order to support linking of lexical entries to their translation equivalents. To help this task we have developed an interactive environment: **TGE**. Several experiments, corresponding to rather "closed" semantic domains, have been developed in order to generate lexical cross-relations between English and Spanish.

Keywords

Lexicons, electronic dictionaries, machine translation.

1 Introduction

Recently, several approaches have been made to extend lexical unification-based formalisms to deal with multilinguistic phenomena in order to be used in applications such as Machine Translation [7].

Within Acquilex II[3] Project, a unification framework based on typed feature structures [4] was developed, the **LKB** (Lexical Knowledge Base), in order to represent conceptual units corresponding to lexical senses, lexical and phrasal rules, multilingual relationships, etc.

This paper describes the enhancements made, to the **LKB** system [6], in order to support linking of lexical entries to their translation equivalents. The organisation of the paper is as follows: Section 2 presents the motivations and formalisation of tlinks (for "translation links"). Section 3 deals with **TGE** (Tlinks Generation Environment), the way we propose to help in constructing lexical linkages semi-automatically from LKB data and bilingual dictionaries [13], [8], loaded in the **LDB** (Lexical Data Base) environment [5]. Section 4 shows the use of TGE within **SEISD**[4] [1] (Sistema de extracción de Información Semántica de Diccionarios). In section 5 some experimental results are presented. Finally in section 6 we present our conclusions and further lines of research.

2 Tlinks

The initial assumption was that the basic units for defining lexical translation equivalence should be the lexical entries in the monolingual LKBs, which should, in general, correspond to word senses in the dictionary. Although in the simplest cases we can consider the lexical entries themselves as translation equivalent, in general more complex cases occur corresponding to lexical gaps, differences in morphologic or lexical features, specificity, etc. [11].

We will therefore represent the relationships between words in terms of tlinks. The tlink mechanism is general enough to allow the monolingual information to be augmented with translation specific information, in a variety of ways. We will first describe the tlink mechanism in the LKB and then outline how some of these more complex equivalences can be represented.

The LKB formalism uses a typed feature structure (FS) system for representing lexical knowledge. So we can define tlinks in terms of relations between FSs. Lexical (or phrasal) transformations in both source and target languages[5] are a desirable capability so that we can state that a tlink is essentially a relationship between two rules (of the sort already defined in the LKB) where the rule inputs have been instantiated by the representations of the word senses to be linked.

As shown in fig 1, *furniture* can be encoded as translation equivalent to the plural *muebles* by specifying that the named rule *plural* has to be applied to the base sense in Spanish. As any other LKB object a tlink can be represented as a feature structure, as shown in fig 2. The type system mechanism, in the LKB, allows further

[1] *This researcher has been suported by a grant of the Departament d'Ensenyament of Generalitat de Catalunya, 91-DOCG-1491.*

[2] *This researcher has been suported by a grant of the Ministerio de Educación y Ciencia, 92-BOE-16392.*

[3] *AcquilexII EC Esprit project BRA 7315.*

[4] *SEISD is an interactive environment built within Acquilex project in order to help in constructing the LKB entries from the LDB sources.*

[5] *in fact tlinks are undirected relations.*

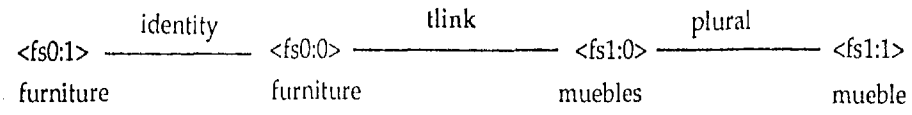

Figure 1: A tlink between furniture and muebles.

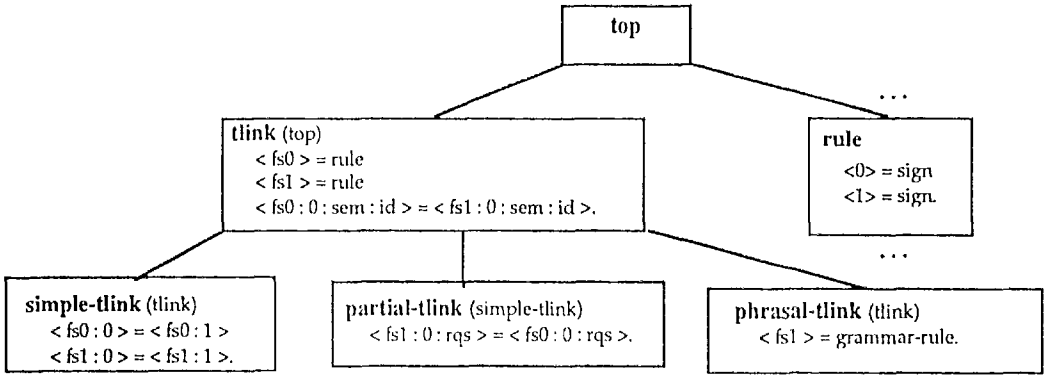

Figure 2: partial view of our tlink type hierarchy.

refinement and differentiation of tlink classes in several ways. A **simple-tlink** is applicable whenever two lexical entries which denote single place predicates (nouns, etc.) arestraightforwardly an equivalent translation, without any previous transformation. Thus, assuming that the LDOCE [9] sense absinth_L_0_1 is translation equivalent to the VOX [12] sense absenta_X_I_1, we will have the next tlink:

```
simple_tlink
< fs0 : 1 > == absinth_L_0_1
< fs1 : 1 > == absenta_X_I_1.
```

The "syntactically sugared" version, which appears in tlink files, is:

```
absinth_L_0_1 / absenta_X_I_1 :
simple-tlink.
```

A **partial tlink** is applicable when we want to transfer the qualia structure from one sense to another, and a **phrasal tlink** is necessary when we need to describe a single word equivalent translation to a phrase [10].

3 TGE: Tlinks Generation Environment

The establisment of tlinks can be obviously performed manually, but the multiplicity of possible cases and the existence of several Knowledge Sources (such as bilingual dictionaries, monolingual LDBs, or a multilingual LKB) allows and motivates the (partial) automatization of the process. To help in performing such a task we have developed an interactive environment: TGE.

TGE has been implemented using a Production Rules approach. This approach was already used within the SEISD environment and was mainly motivated by the need of providing a flexible and open way of defining tlink formation mechanisms. The core of TGE is **PRE** (production rules environment), a rule-oriented general purpose interpreter [2]. PRE follows the philosophy of most **Production Rules Systems** [3] but is deeply adapted to Natural Language applications. **PRE** offers a powerful (according to both expressiveness and performance) rule application mechanism and provides the possibilities of defining higher level mechanisms, such as rulesets (allowing inheritance capabilities) and of choice among control strategies, either user-defined or provided by the system. Consider the following example:

```
(rule rule-1-all
 ruleset all
 control forever
 priority 1
 (translation-in ^trans-records  (?translation *rest))
 ->
 (modify 1 ^trans-records (*rest))
 (create translation
        ^trans-psorts nil
        ^trans-record ?translation
        ^tlink-type nil ^checked nil))
```

In this rule the pattern-condition is the occurrence of an object named **translation-in** in the Working Memory. This object must in turn contain a **^trans-records** attribute whose value will be matched against the pattern **(?translation *rest)**. If the matching succeeds then

translation will be unified with the first element of the list and **rest** with the remainder elements. The action part of the rule consists of two actions. The former is the modification of **translation-in**, popping its first element, and the latter performs the creation of another object, named **translation**. **Rule-1-all** rule is applied until all the objects named "translation-in" have emptied the list contained in their slot ^trans-records.

4 Using TGE for generating Tlinks

The TGE may be considered a toolbox and thus, it doesn't impose a fixed methodological strategy. Whatever methodology we follow, several decisions must be taken: the kind of control we need, the rulesets to be designed, the rules belonging to each ruleset, the relative priority assigned to each rule and so on.

As regards the control strategy, one of the following four alternatives may be chosen for each source entry (see [2] or [10] for futher details):

• **All**, which executes all the rulesets. From the proposed tlinks, finally the user chooses the correct ones.

• **Collect**, which executes the rulesets one at a time and provides the results to the user (for selection of the correct ones) every time a ruleset succeeds.

• **One-by-one**, which orderly executes the rulesets and stops as soon as one of them succeeds.

• **Select**, which only executes the rulesets that the user chooses.

An initial set of modules was designed according to the typology of tlinks presented so far. It included four sorts of tlinks that showed distinct conceptual correspondences between both languages. A more in-depth study of English-Spanish mismatches [11] might lead to an enrichment of the typology, and consequently to a need for extending the extant modules.

Up to now seven modules, each one implemented as a ruleset, have been developed. Each of them generates one out of the three kinds of tlinks stated above. Each module follows a different strategy to guess a possible tlink, taking account of the three accessible knowledge sources.

• **Simple Tlink Module**, this is the case when there exists a direct translation of the source entry in the bilingual dictionary. Example:

```
absenta_x_i_1    --> absenta          LKB source entry
absenta          --> absinth          bilingual dictionary
absinth          --> absinth_L_0_1    LKB target entry
==>
absenta_x_i_1 / absinth_L_0_1:
SIMPLE-TLINK.
```

"absenta" is translated in the bilingual dictionary by "absinth", ABSINTH_L_0_1 is a valid lexical entry of the target lexicon, and therefore a SIMPLE-TLINK connecting both entries is created.

• **Orthographic Tlink Module**, this case occurs when in both languages the same word with exactly the same spelling is used. Therefore, no bilingual dictionary is needed.

• **Compound Tlink Module**, this is the case when the corresponding entry in the target lexicon is a compound one, being the target lexical entry made up of the concatenation of the two English words that appear in the bilingual entry.

• **Phrasal Noun Tlink Module**, this case takes place whenever the translation is the concatenation of two other nouns; for example, the Spanish nouns for trees often correspond to two nouns in English, (like limonero - lemon tree, melocotonero - peach tree, etc.). More complex cases can be recovered by using different grammar rules (also implemented within the LKB formalism).

• **Parent Tlink Module**, this is the case of those very specific terms in the source lexicon which are not treated in the bilingual dictionary, but whose hyperonyms in the taxonomy have a clear translation that can generate a partial tlink.

• **Grandparent Tlink Module**, this is a very similar case to the previous one, in which the source word's grandparent is used to produce the partial tlink.

• **General Tlink Module**, this is the case when the translation appearing in the bilingual dictionary is composed of more than one word. Normally these explanations are made up as definitions in the form of a genus, plus some modifiers. A tlink connecting the source entry and the genus appearing in the definition must be created.

We will illustrate the tlink generation process with an example of an entry for which a number of different tlinks have been generated, namely *batido_X_I_5*. In figure 3 where *batido_X_I_5* appears with the tlink options, we had selected the option *all* and subsequently, all the possible tlinks have been suggested by the system. However the TGE allows for other selection criteria. As we can see in figure 3, five tlinks are suggested by the system for this particular example:

1) The first option is not a correct one. Among the various translations given for the source LKB entry *batido_X_I_5* the adjective *shot* appears. Because another syntactic realisation of *shot* is that of a noun denoting a drinkable thing therefore it is included in the target subset.

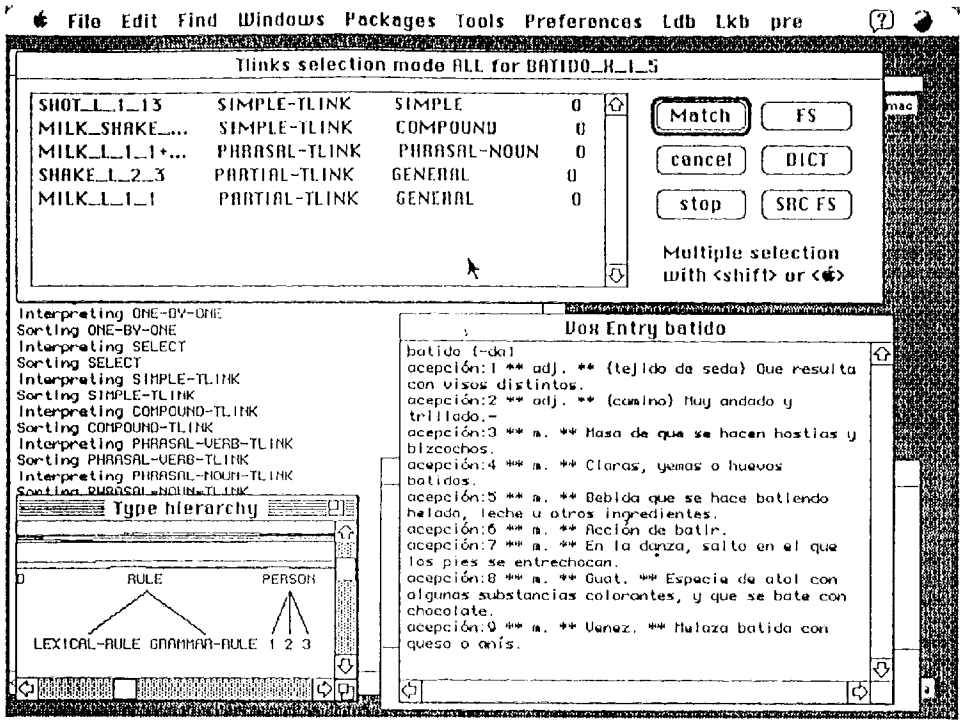

Figure 3: Options for creation of tlinks .

2) The second is a simple-tlink type linking *batido_X_1_5* with the target LKB entry *milk_shake_L_0_0*. In this case we have an example of the application of the compound-tlink-ruleset.

3) The third is a phrasal-tlink type, linking *batido_X_1_5* with the target LKB entries *milk_L_1_1* and *shake_L_2_3* composed by the + sign. This is an example of the application of the phrasal-noun-tlink-ruleset.

4) Both the fourth and fifth are partial-tlink-types, linking *batido_X_1_5* with the target LKB entries *shake_L_2_3* and *milk_L_1_1* respectively. This is an example of the application of the general-tlink-ruleset.

5 Results

Several experiments corresponding to rather "closed" and narrow semantic domains have been performed. We discuss next those corresponding to "drinks" [10].

The Spanish taxonomy of drink-nouns, extracted from VOX dictionary, consists of 235 noun senses and has 5 levels. The English taxonomy of drink-nouns, extracted from LDOCE, consists of 192 noun senses. Some of the obtained results are the following:

• While translating from Spanish to English, 223 out of 235 drink-nouns have been linked by means of different and often more than one tlinks (95 %). However, only 52 English nouns have been linked with Spanish nouns (27%). Out of these 223 drink-nouns mentioned above, 210 have been linked by using (mainly) the bilingual dictionary as a translation resource while the rest, 13 of them, have been linked by means of the orthographic-tlink ruleset and consequently, the gap of the bilingual dictionary has finally been bridged, for in both languages the same word with exactly the same spelling is used. For example, *chartreuse_X_1_1* and *chartreuse_L_1_0*, *sherry_X_1_1* and *sherry_L_0_0* , etc.

• 74 out of 235 source LKB entries for drink-nouns are also bilingual entries (31,5%). Consequently, 161 source LKB entries have no corresponding bilingual entries (68,5%). This big gap in the bilingual dictionary is due to the fact that the one used, VOX/Harrap's, is a very basic one, and as such it only contains 32,463 senses. In contrast the VOX monolingual Spanish dictionary covers a total of 143,700 senses.

• 30 out of the translations of the 74 source LKB entries which were found in the bilingual dictionary are also target LKB entries. Consequently, the translations of 44 bilingual entries have no corresponding target LKB entries.

• 13 out of 161 source LKB entries are also target LKB

entries (8 %).

• For most entries, more than one tlink type has been extracted. The total number of tlinks which have been generated and selected for the taxonomy of *bebida* _X_I_3 *(drink)* with the explained software is 372 tlinks. Next we show the different tlinks generated by each ruleset and the amount of lexical entries of each language involved.

		(a)	(b)	(c)
simple-tlinks	(14,5%)	55		
by simple-tlink-ruleset		41	26	31
by compound-tlink-ruleset		1	1	1
by orthographic-tlink-ruleset		13	13	13
phrasal-tlinks	(0.5 %)	2		
by phrasal-noun-tlink-ruleset		2	1	3
partial-tlinks	(85 %)	320		
by parent-tlink-ruleset		268	149	15
by grandparent-tlink-ruleset		44	30	10
by general-tlink-ruleset		8	7	6

(a) Total Number of Tlinks
(b) Spanish entries
(c) English entries.

6 Conclusions

In this paper we have presented TGE, an environment designed and built in order to help in the recovery of cross-linguistic relations. We have reported and described results of an experiment for automatically extracting the relations of equivalence for Spanish and English drink-nouns by using the TGE software. The resulting process is semi-automatic, whilst the tlink generation is performed automatically, the selection of the desired tlinks is done manually.

All the tlink-rulesets have worked satisfactorily, therefore resulting in a considerable part of the subsets being linked (95% of the source lexicon). However these PRE tlink-rulesets have only been tested over limited subsets of specific semantic fields. Its actual potential will be proven in a later stage, once its application to larger and less restricted sets of word senses takes place, including also categories which are not considered to be nouns.

7. References

[1] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M., Verdejo F., *SEISD: An environment for extraction of Semantic Information from on-line dictionaries.* Proceedings of 3th Conference on Applied Natural Language Processing. Trento. Italy. 1992.

[2] Ageno A., Ribas F., Rigau G., Rodríguez H., Verdejo F., *TGE: Tlink Generation Environment.* Esprit BRA-7315 Acquilex II Working Paper. 1993.

[3] Brownston L., Farrell R., Kant, E., Martin N., *Programming Expert Systems in OPS5.* Addison-Wesley. 1986.

[4] Carpenter B, The Logic of Typed Feature Structures, Cambridge University Press, Cambridge, England, 1992.

[5] Carroll J. *Lexical Data Base System. User Manual.* Computer Laboratory, University of Cambridge. 1990.

[6] Copestake•A., *The Acquilex LKB: represention issues in semi-atomatic acquisition of large lexicons.* Proceedings of 3th Conference on Applied Natural Language Processing. Trento. Italy. 1992.

[7] Copestake, A., Jones B., Sanfilippo A., Rodriguez H., Vossen P., *Multilingual Lexical Representation.* Esprit BRA-3030 Acquilex Working Paper n°38. 1992.

[8] Hastings A., Rigau G., Soler C., Tuells A. *Loading a bilingual dictionary into the LDB.* Esprit BRA-7315 Acquilex II Working Paper. 1993.

[9] Procter, P. et al. (eds). Longman Dictionary of Contemporary English. Longman, Harlow and London. 1987.

[10] Samiotou, Anna, *Performance of cross-linguistic equivalence relations: A lexicon-based approach.* Msc. Dissertation. UMIST. 1993.

[11] Soler, C., *Dealing with Spanish-English/ English-Spanish mismatches.* Esprit BRA 7315 Acquilex II Working Paper. 1993.

[12] Diccionario General Ilustrado de la Lengua Española VOX. Ed. Biblograf S.A. Barcelona, 1987.

[13] VOX Harrap's Diccionario esencial Inglés-Español, Español-Inglés. Segunda Edición. Biblograf S.A. Barcelona, 1992.