Translation Equivalence via Lexicon: A Study on Tlinks

Anna Samiotou
Irene Castellón
Francesc Ribas
German Rigau

TRANSLATION EQUIVALENCE VIA LEXICON: A STUDY ON TLINKS.

Samiotou A., Castellon I., Ribas F., Rigau G.

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

The first section of the present paper provides a study and suggests an extension of the tlink mechanism. The second section represents the practical part of the paper, namely the performance of cross-linguistic equivalence relations within the Acquilex II framework. It explains the process of the tlink generation experiment and it extensively discusses the results by giving a lot of illustrative examples.

1 Study and extension of tlinks.

During the process of mapping the monolingual source and target lexicons we are dealing with two steps (see figure 1):

- mapping an entry from the monolingual source lexicon to an entry in a bilingual dictionary, and

- mapping the translation(s) of the source entry, given in the bilingual dictionary, to the entries in the monolingual target lexicon.

source lexicon ↔ bilingual dictionary ↔ target lexicon

Figure 1: Graphic representation of mapping lexicons

From the theoretical point of view, no global study of tlinks [Copestake A. & Sanfilippo A., 1993] has been made so far, that is there is not any complete plan available that includes all the possible cases which could appear during the mapping process. It would be interesting and very useful to see what all the possible cases that we could come across are, when having a source lexical entry, we want to get a translation equivalent by means of a tlink in the target language. Interesting, because this plan would allow us to gain a global view of the cases that we actually have to deal with. Useful, because we could set a tlink framework and check which of these cases are dealt with in the present system and how succesful their treatment is; whether some changes, due to different classifications and groupings should be made in the whole rule-production system (PRE) as well as in its individual tlink modules; finally, in order to be aware of what other tlink modules have to be made. Following a theoretical framework, therefore, would enable us to improve and extend the tlink mechanism with a steady pace, in the most efficient possible way.

Due to the present situation with respect to tlinks, the different mapping and linking modules are constructed occasionally, and not enough attention is paid to the way they
combine with each other or to what other extenensions can be made to these modules.

For example, a case which was not at all encountered until very recently, was when the input source entry was not found as an entry in the bilingual dictionary, but its translation equivalent existed in the target dictionary (its existence is identified by the lexicographer's/user's knowledge). This was due to the fact that the (current) mapping program essentially created tlinks (and it still does) by means of consulting, mainly, the bilingual dictionary as well as the monolingual dictionaries. A solution has been given to a subcase of the above case, when the source entry happens to be an entry in the target dictionary as well (that is, with exactly the same spelling), by means of the orthographic-tlink module. If this module succeeds, a simple-tlink is generated.

There are cases, however, where the orthographic-tlink module would be expected to succeed, but it fails instead; this often happens with foreign words which are incorporated in the vocabulary of a language and they are spelled according to the spelling rules of this language. So the spelling of a word is not exactly the same in two languages due to the spelling rules of a each language. For example, consider the word anis in English; the corresponding word in Spanish happens to be the same, with a slight, however, difference which comes from the addition of an accent on the "i". So the Spanish word is anís, due to a particular spelling rule which defines that all the words that end in s, z and ñ which are also accentuated to the last syllable, take an accent on the last vocal. Other examples are vodka-vodca, whisky-guïsqui, and so on so forth. By grouping all these rules together, it would be possible to construct some new rulesets and LISP functions to provide a solution for the above cases as well.

Another possibility for the source lexical entry, before start looking for its ancestors (if it is not found in the bilingual dictionary as such), would be to check its morphology, just in case it happens to be the result of a derivational process, or a compound, or even two words concatenated by a hyphen and so on. By means of adequate algorithms or lexical/grammar rule applications we could try to get hold of the stem of the source word or kernel in the case of the synthetic compound and see whether it exists as such in the bilingual dictionary or not. In the positive case, we proceed to the next step, that is the mapping to the target dictionary and so on.

The continuation of the procedure may involve further complicated cases, as for example, correspondence between mass and count nouns (furniture - muebles), etc. (see also [Sanfilippo A. et al., 1992]). If no direct tlink is generated for the input source entry (because of gaps in the bilingual or/and the target monolingual dictionary), the partial-link based on the translation of its ancestors, that is its parent and grandparent via taxonomy, is considered.
Before going on, something else that has to be pointed out is that, in the current mapping program it is always assumed that the source entry is a single word; however this is not always the case because it is perfectly possible to have more than one words as an entry in a dictionary, although it does not happen very often. A number or such examples are found in English language dictionaries, as malt milk, milk shake, home brew (beer) and so on.

At this point, we would like to discuss some enhancements and changes in the tlink mechanism (see also [Samiotou A., (forthcoming)]¹). First of all, we will display and reason the new hierarchy of tlink-types; immediately after, we will discuss the enhancements in the partial-tlink modules as well as in the phrasal-tlink modules.

The tlink-type hierarchy we propose is the following (this is how it appears in the tlink-type file):

```
tlink (top)
<fs0> = rule
<fs1> = rule
<fs0:0 : sem : ind> = <fs1:0 : sem : ind>.

simple-tlink (tlink)
<fs0:0> = <fs0:1>
<fs1:0> = <fs1:1>.

phrasal-tlink (tlink)
<fs0:0> = <fs0:1>
<fs1> = grammar-rule.

partial-tlink (tlink)
<fs0:0> = <fs0:1>
<fs1:1 : sem> = <fs1:0 : sem>
<fs1:1 : orth> = <fs1:0 : orth>
<fs1:1 : cat> = <fs1:0 : cat>
<fs1:1 : sense-id> = <fs1:0 : sense-id>
<fs0:0 : rqs> = <fs1:0 : rqs>.

partial-phrasal-tlink (phrasal-tlink partial-tlink).
```

We can also represent the new tlink-type hierarchy in a tree form, as follows:

---
¹ This MSc dissertation is focussed on tlinks; it has been carried out at UPC, in Barcelona, and was submitted at the University of Manchester, UK.

Translation Equivalence Via Lexicon: A Study on Tlinks.
The instances of the different tlink-types are the following: of the simple-tlink-type, the simple-tlink, the orthographic-tlink and the compound-tlink; of the phrasal-tlink-type, the phrasal-noun-tlink and the phrasal-verb-tlink; and of the partial-tlink-type, the parent-tlink, the grandparent-tlink and the general-tlink. The partial-phrasal-tlink type is a new type which is generated when we link the source word with the phrasal translation of its ancestor (see below) and it has to fullfil the constraints of both the partial and phrasal tlink-types.

With respect to the phrasal-tlink, the current idea of dealing with phrasal translation is to analyse the phrase, find the kernel and link the source word with the translation of the kernel, by means of a phrasal-tlink. Given the fact that there is no efficient analyser for English available here in Barcelona, the phrasal-tlink discussed previously cannot be applied as such. The (general) phrasal-tlink-type was therefore created to somehow offer a solution. This, however, was considered as a subtype of the simple-tlink-type, but this is not correct because there is no direct equivalence between the linked words, as the target word with which we link the source word shares only partially semantic or/and syntactic features. In this case, the relation between the two words is partial; general indicates the linking of a more specific to a more general word and so, this should involve transfer of the missing information. Consequently, it seems more reasonable to consider this relation as an instance of the partial-tlink-type and call it, henceforth, general-tlink.

If the translation of the input source word is a phrase in the bilingual dictionary, the general-tlink module will generate partial-tlinks with all the words of the phrase that are found in the target lexicon. The fact that it is now classified as a partial-tlink-type provides a further filter for the generated tlinks. In general, when we ask for the
expanded tlink-types, the generated tlinks have to fulfill all their corresponding constraints. One of the constraints of the partial-tlink-type is the following:

\[ < fs1 : 1 : cat > = < fs0 : 0 : cat > \]

which means that the category of the linked words must be the same.

Let us take the example of the Spanish word *amontillado* whose translation often given in English is *pale dry sherry*. As we work with limited subsets of the semantic fields, only the word *sherry* is found and, thus, it is the only one which is linked by means of the partial-tlink with the source word *amontillado*. Assuming that the word *pale* was also found, we would also get as a result of the generation process the following (despite of the fact that this is not a relation between a more specific and a more general word):

*amontillado*\_\_X\_1\_1 \hspace{1em} / \hspace{1em} *pale*\_\_L\_1\_1

partial-tlink

However, at the moment of the expansion of the tlink-type, when the unifications of features take place and the type constraints are applied, the partial-tlink between *amontillado* and *pale* will not be accepted because they clash in category as the former is a noun whereas the latter an adjective.

On the other hand, we should take notice of the group of the partial-tlinks that are generated by means of the general-tlink-module for a specific source word; if they are more that one and, in addition, no phrasal-tlink is given for it, this means that the adequate phrasal-tlink-module is missing. Consequently we get a hint that we should think about how to deal with that case and, further, construct the ruleset, make the necessary modifications in the PRE and create some new LISP functions if needed. For example, consider the Spanish word *novillo* and its English translation *young bull* and suppose that the system generated the following tlinks:

*novillo*\_\_X\_1\_1 \hspace{1em} / \hspace{1em} *bull*\_\_L\_1\_1

partial-tlink

*novillo*\_\_X\_1\_1 \hspace{1em} / \hspace{1em} *young*\_\_L\_1\_1

partial-tlink

(The program informs us that the general-tlink-module has been applied). From these results, we realise that there is no phrasal-tlink-module available to deal with this case, that is a module that combines a noun with an adjective by means of a grammar rule.

With respect to the phrasal translations, apart from providing for a partial-tlink by
means of the general-tlink-module discussed above, the system permits us (at the moment) to deal with the following two cases:

- a verb whose translation is another verb followed by a particle, for example, the Spanish *aupar* which is translated in English to *lift up*. This case was already dealt with in the former system; however, now, the phrasal-verb-tlink is only an instance of the phrasal-tlink-type and not a type itself, as it merely involves the application of the *verb_part_application* grammar rule which is an instantiation of the grammar-rule-type which appears in the definition of the phrasal-tlink-type. As a result, the generated tlink is a phrasal-tlink.

- a noun whose 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. This two English nouns, normally form separate lexical entries in the lexicon. This is a new case which combines the two (existing in the target lexicon) nouns by means of the *forward_application* grammar rule (this instance of the grammar-rule-type is applied because in English, normally, the head noun is on the right hand side, and so the modifier noun looks for it towards the right, that is, it follows a forward-direction). The result we get is the following:

`limonero_X_I_1 / lemon_L_I_1+ tree_L_I_1`  
phrasal-tlink

The new ruleset consists of four rules and is the following:

```
(rule rule-1-phrasal-noun-tlink
 ruleset phrasal-noun-tlink
 priority 1
 control one
 (not tlink-type)
 ->
 (create tlink-type ^type phrasal-tlink ^tlink-ruleset phrasal-noun))

(rule rule-2-phrasal-noun-tlink
 ruleset phrasal-noun-tlink
 priority 2
 control forever
 (tlink-type ^type ?type ^tlink-ruleset ?tlink-ruleset)
 (translation-bill ^trans-record ?translation ^viewed nil)
 ->
 (modify 2 ^trans-record ?translation ^viewed t)
 (?translation-psorts :=
 (mapcar
 `'(lambda (psort-list)
 (compose-psorts
 (filter-psorts ?translation psort-list))))
```

*Translation Equivalence Via Lexicon: A Study on Tlinks.*
(create translation
  ^trans-psorts ?translation-psorts
  ^trans-record ?translation
  ^tlink-type ?type
  ^tlink-ruleset ?tlink-ruleset
  ^checked nil))

(rule rule-3- phrasal-noun- tlink
  ruleset phrasal-noun-tlink
  priority 3
  control one
  (tlink-type)
  ->
  (delete 1))

(rule rule-4-phrasal-noun-tlink
  ruleset phrasal-noun-tlink
  priority 4
  control forever
  (translation-bil ^trans-record ?translation ^viewed t)
  ->
  (modify 1 ^trans-record ?translation ^viewed nil)

The first rule creates an object called tlink-type (the phrasal-tlink, in particular), in case such object does not exist; If such an object exists in the WM, that implies that the ruleset has been called either by the parent-tlink ruleset or the grandparent-tlink ruleset. The second rule's condition is the existence of an object of type translation-bil in the WM with the stated pattern. The object translation-in is created in the top ruleset and the object translation-bil in the all, collect and one-by-one rulesets (see below). If this condition is satisfied, the action, for the program, is to modify this object by changing its last attribute's value to 't' so that the same translation of the source word is not reconsidered. Next, the attribute translation-record, kept in ?translation, is taken and its orthography is captured (ie, its translation definition) by means of the LISP function translation-record-target-orth. If it is composed of two words which can be combined by the phrasal-noun LISP function in a list of concatenations of two psorts, then once these two nouns are obtained they are filtered (by the filter-psorts LISP function) in order to get hold of their psort-name, and the results are concatenated with the + sign. The final results are kept in the value ?translation-psorts. Finally, a translation object is created with the values obtained. The third rule deletes the tlink-type object so that the ruleset can be used again with another (or even the same) tlink-type object. The fourth rule modifies the translation-bil object so that the ruleset can be applied recursively, until no more translations of the source entry are left.

The creation of the two objects we referred to previously is made (at least) in the following
rules:

(rule rule-6-top
  ruleset top
  control one
  priority 3
  (source-word ^orth ?orth ^psort ?psort)
  (not execution-mode ^option stop)
  ->
  (?trans-records := (consult-bilingual ?orth))
  (create translation-in ^trans-records ?trans-records)
  (create translation-out ^targets nil))

(rule rule-1-all
  ruleset all
  control forever
  priority 1
  (translation-in ^trans-records (?translation *rest))
  ->
  (modify 1 ^trans-records (*rest))
  (create translation-bil ^trans-record ?translation ^viewed nil))

In a similar way, we could extend the phrasal-noun-module for more than two nouns, or construct new rulesets that would, for example, combine adverbs with adjectives, noun with adjectives and so on. The latter one, in particular, would sometimes involve further complications; consider examples like patata frita (fried potato) where the adjective frita is, on the one hand, filling one of the features associated to the type of the head patata, and on the other it associates to a certain type which can support being fried.

Dealing with such relationships between nouns and adjectives implies a definition of the types in the LKB [Martí M.A. & Soler C., 1993].

With respect to the partial-tlink-module, whenever it was applied until now, it only dealt with the case where the ancestor(s) of the source word had a single word translation. That is, the module only use to call the simple-tlink-module. By making the adequate modifications and additions to all the tlink modules, it is now possible to deal with the case where the ancestor(s) of the source word has a phrasal translation. As a result, now we get far more tlinks than before, as the tlink mechanism can capture more cases than it used to. For example, suppose we have the following taxonomy:
The generated tlklinks by the suggested version of rulesets would be the following:

- C / K : simple-tlink
- C / F+G : partial-phrasal-tlink
  (after application of parent-tlink and phrasal-noun-tlink modules)
- C / F : partial-tlink
  (after application of parent-tlink and general-tlink modules)
- C / G : partial-tlink
  (after application of parent-tlink and general-tlink modules)
- C / D_E : partial-tlink
  (after application of parent-tlink and compound-tlink modules)

whereas before we would only get the first one.

The rule of the parent-tlink-ruleset which calls the other rulesets has now taken the following form:

```
(rule rule-4-parent-tlink
 ruleset parent-tlink
 priority 4
 control forever
 (translation-in-parent)
 (translation-bil)
 ->
 (create tlink-type ^type partial-tlink ^tlink-ruleset parent-simple)
 (apply-ruleset simple-tlink)
 (create tlink-type ^type partial-tlink ^tlink-ruleset parent-compound)
```

Translation Equivalence Via Lexicon: A Study on Tlinks.
That is, if both the objects translation-in-parent and translation-bil exist in the WM, then the tlink-types are created with their respective attributes and values, and immediately after, the corresponding rulesets are applied. Notice that, the second attribute of the object tlink-type is very useful because when a generated partial-tlink (or partial-phrasal-tlink) is displayed on the screen, the value of this attribute indicates the rulesets which have been applied during the process (it is a sort of trace).

To conclude, we regard that the proposed hierarchy provides a better and more natural classification of the tlink-types. Moreover, the enhancements and changes in the rulesets have improved a great deal the tlink generation process as well as its results.

In the next chapter, we will describe an experiment for the semi-automatic generation of tlks between nouns for the semantic domain of bebida (drink).

2 Tlink generation experiment.

2.1 Process.

In order to initiate the tlink generation process, we have to load:

i) the TGE [Ageno A. et al., 1993a, 1993b] which consists of the systems LDB, LKB and PRE.

ii) the LISP functions which are common to all the PRE programs as well as the auxiliary ones that are necessary for the execution of the program PRE tlink-rulesets.

iii) the lexicons we need for our experiment, that is, the Spanish lexicon which contains the taxonomy of drink-nouns (bebida) and the corresponding English lexicon (LKB entries).

iv) the index of these lexicons (if they are not yet indexed, by selecting the option index and index and check from the LKB menu the system indexes them automatically).
v) the tlink-rulesets.

The next step is to select the option map from the LKB menu which, in turn, consists of the following:

(a) define languages 
(b) select source entries 
(c) select bilingual 
(d) perform mapping 

In (a) we define Spanish as source language and English as target language. In (b) the selection of the source entries can be done in various ways, namely, by means of the taxonomy option which gives the entries of the taxonomy, whose top we have to specify, in alphabetical order; or by means of the all entries option which gives all the entries that the loaded source lexicon contains in alphabetical order (the distinction in these two options lies in the fact that if we have more than one taxonomy in the loaded source lexicon, the first option would only give us the entries that correspond to the taxonomy (or taxonomies) we have initially specified, whereas the second option would give us all the source entries. Finally, another way offers the individual option, where we can type in the source entries we wish the system to generate links for. In (c) we load the bilingual Spanish-English VOX Harrap’s essential dictionary in LDB format, which will be used as the main resource for the mapping. In (d) we ask the program to start with the tlink generation process; at this point we create the file where we wish the generated tlinks to be kept and the process gets started. 

A window with the possible tlink options is shown to the user, namely (see figure 4)

• all, which executes all the rulesets 
• collect, which executes the rulesets one at a time and gives the results every time a ruleset succeeds 
• one-by-one, which executes the rulesets in order and stops the process (for the source LKB entry in question) in the first success 
• select, which executes the rulesets that the user chooses 
• cancel, which ignores the shown entry and goes to the next one 
• stop, which stops completely the process for the source lexicon entries 

Translation Equivalence Via Lexicon: A Study on Tlinks.
Figure 4: Options for ruleset execution

We have mainly used the *all* option, which has given me all the possible tlinks for the source entries. We would like to point out that a source entry can have many tlinks with target entries; for example, in the Spanish taxonomy, _coñac_X.L.1_ is descendent of _aguardiente_X.L.1_ which, in turn, is descendent of _bebida_X.L.3_. As it appears in the appendix, _coñac_X.L.1_ is linked by means of a simple-tlink with _cognac_L.0.0_; however, other tlinks have also been generated for this source entry, namely, four partial-tlinks with _liquor.L.0.1_, _liquor.L.0.2_, _brandy.L.1.1_, _brandy.L.1.2_, generated by the parent-tlink ruleset as well as another partial-tlink with _drink.L.2.1_, generated by the grandparent-tlink ruleset. These five partial-tlinks have not been included in the appendix, as we have only included the most direct ones. The partial-tlinks which appear in the appendix indicate at the same time that there are no more direct tlinks for the particular source entries.

We will illustrate the tlink generation process with a (full) example of an entry for which a number of different tlinks have been generated, namely _batido_X.L.5_. In the screen where _batido_X.L.5_ appeared with the tlink options, we have selected the option *all*, consequently, all the possible tlinks will be suggested by the system.

In most cases, two (successing) screens with generated tlinks are expected to be shown to the
- the first screen corresponds to the tlinks generated by using the translation(s) of the source LKB entry given in the bilingual dictionary (if there are any) and which are, in turn, matched to target LKB entries

- the second screen corresponds to the tlinks that deal with the bilingual gaps.

Both screens have four columns; the first shows the linked target LKB entry, the second gives the type of the generated tlink, the third shows the rulesets that have been applied in order to generate this tlink-type and, finally, the fourth gives the weights of the matchings.

![Figure 5: Options for creation of tlinks](Image)

It must be pointed out that the LKB entry `milk_L_1_1` is not included in the target lexicon because it does not belong to the `drink_L_2_1` taxonomy; however, we have only temporarily included it, for the sake of this example, because it gives a number of interesting results which are worth discussing.
As we can see in the first screen, five tlinks are suggested by the system.

1) The first one is discussed further down.

2) The second, is a simple-tlink-type linking batido_X_I_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_I_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_I_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.

Besides the tlink-window, there is also another window (which we have asked for), showing the format of the monolingual LDB entry.

The next step is to select the desired generated tlink(s) (which is the second simple-tlink in this case), and press the option match. Following, the second screen will appear, suggesting a partial-tlink generated by means of the parent-tlink-ruleset, which links batido_X_I_5 with the target LKB entry drink_L_2_1.
Figure 6: More options for tlink creation

This is an example of the application of the parent-tlink-ruleset. As we have already found and selected a simple-tlink, we do not need to select the last one because it is far more indirect. After the option `match`, a third screen appears, containing only the selected tlinks with their respective matching scores.

Besides the tlink-window, there is also another window (which we have asked for), showing the format of the bilingual LDB entry.
At this point the final tlink selection takes place and the final results are kept in the tlink file, which is the following, in our example:

batido_X_I_5 / milk_shake_L_0_0 :
simple-tlink.

2.2 Results.

The taxonomy is as complete as possible, according to the Spanish monolingual VOX dictionary. That is, the taxonomy has been constructed taking information exclusively from the above dictionary. Its construction is based on the definitions of the drink-denoting nouns; only those that are defined as *bebida* (drink) have been included as direct taxonomic children of the top *bebida_X_I_3*. So, the word *bebida* was taken as the genus of the definition. Further, the taxonomy also contains the descendents of the children of its top. As a result, the taxonomy does not form the whole subset of the semantic field in question. There are, for example, words denoting drinkable things, like most kinds of cocktails (e.g. daiquiri ) which are not defined as taxonomic children of *bebida* in the
VOX dictionary. Another group of drink-nouns which has not been included (even if they are defined as such) are the words that are only used in Latin America (but not in Spain); the decision of not including them in the taxonomies was taken because otherwise the taxonomies would have resulted very big.

The Spanish taxonomy of drink-nouns consists of 251 nouns (i.e. senses), and has 5 levels. To start with, we will discard a number of them, 16 in total, because they do not really belong to this taxonomy. For example, I discarded chicha and pulque as Mexican words, cuba (only its collocation with ~libre is a drink and not on its own), cubalibre (as child of cuba; in any case it is a cocktail and as such it is not defined as a drink at all), etc. As a result, the number of source lexical entries that I will treat is 235 (i.e. senses) nouns in total. The English taxonomy of drink-nouns consists of 192 nouns (i.e. senses). From these, only 52 nouns have been linked with Spanish nouns (27%).

Going from Spanish to English, 223 out of 235 drink-nouns have been linked by means of different, often more than one, tlinks (95%). 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, that is 13, have been linked by means of the orthographic-tlink ruleset, and, consequently, the gap of the bilingual has been bridged in the end, because in both languages the same word with exactly the same spelling is used. For example, chartreuse_X_L_1 and chartreuse_L_L_0 sherry_X_L_1 and sherry_L_0_0 and so on (see appendix).

Further, some more results are the following:

<table>
<thead>
<tr>
<th>source entries</th>
<th>VOX/Harrap's entries</th>
<th>target entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>235</td>
<td>74</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 (due to polysemy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13 (identical spelling)</td>
</tr>
</tbody>
</table>

- 74 out of 235 source LKB entries drink-nouns are also bilingual entries (31,5%). Consequently, 161 source LKB entries have no correspondent bilingual entries (68,5%). This big gap in the bilingual is due to the fact that the bilingual VOX/Harrap's dictionary we use is an essential one, and as such it only contains 16,463 entries and has an average of 1,99 senses per entry, that is, the total number of senses is 32,463. By contrast, the VOX monolingual Spanish dictionary has a total of 89,793 entries and an average of 1,6 senses per entry, that is the total number of senses is 143,700.

- 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 54 bilingual entries have no correspondent target LKB entries.

- 13 out of 161 source LKB entries are also target LKB entries (8 %)

- the total number of target polysemous drink nouns is 11; 8 of them belong to the group of the linked target words which correspond to 18 target LKB entries. In the Spanish lexicon there are only 2 polysemous nouns, which correspond to the following LKB source entries, \textit{bebedizo}_X\_1\_2 and \textit{bebedizo}_X\_1\_3 as well as \textit{filtro}_X\_II\_1 and \textit{filtro}_X\_II\_2.

It should be pointed out that during the mapping from:

- the source lexicon to the bilingual dictionary, and
- the bilingual dictionary to the target lexicon,

we cannot map senses because lexical entries (of the bilingual dictionary) as well as the translations given in the bilingual dictionary are not distinguished for their sense. In most cases, the various candidates given by the bilingual dictionary are distinguishable on the basis of their lexical semantic type alone, and this is due to the fact that we are dealing with limited subsets of a particular semantic field. For example, in the translations given in the bilingual dictionary for the source LKB entry \textit{matarratas}_X\_I\_2 are the following: (1) \textit{rotgut.} and (2) \textit{rat poison.} The system captures both of them and tries to match them with target LKB entries. For the first given translation, the target entry \textit{rotgut}_L\_0\_0 is found, and, consequently, the following simple-tlink is generated and proposed to the user:

\textit{matarratas}_X\_I\_2 \ / \ \textit{rotgut}_L\_0\_0

simple-tlink

Of course, this tlink has been accepted (see appendix).

For the second given translation, the target entry \textit{poison}_L\_1\_3 is found and, as a result, the following partial-tlink is generated (by means of the general-tlink-ruleset), and proposed to the user:

\textit{matarratas}_X\_I\_2 \ / \ \textit{poison}_L\_1\_3

partial-tlink

Obviously, this tlink has not been accepted, as the particular sense given in the bilingual dictionary does not belong to the semantic domain of drink. The case here is that, despite of the fact that the translation given in the bilingual is irrelevant to the semantic field in
question, it is all the same matched a target LKB entry, generating thus, a wrong tlink. There are three more similar cases observed in the present experiment, which we list below:

a) among the various translations given for the source LKB entry batido_X_L_5 the adjective shot appears; another syntactic realization of shot is that of a noun, denoting a drinkable thing and as such it is included in the target subset. Consequently, the following wrong tlink is generated by the system:

\[
\text{batido}_X\_5 / \text{shot}_L\_1\_13
\]

\text{simple-tlink}

This tlink is wrong for two reasons:

- firstly, because it is the result of a (scaled) mapping between words belonging to two different syntactic categories. To make it more clear, the noun batido is taken as an adjective in the bilingual dictionary, giving the adjective shot as its translation which is, in turn, mapped to the noun shot in the target lexicon. This happens due to the lack of a control process with respect to the syntactic categories when mapping two words

- secondly, because the translation of the Spanish noun batido is not the English noun shot.

b) the reasoning is exactly the same for the next example, which is:

\[
\text{amargo}_X\_1\_7 / \text{bitter}_L\_3\_0
\]

\text{simple-tlink}

c) the only difference in the following example is that in the (scaled) mapping the nouns are considered as verbs due to the bilingual information:

\[
\text{beber}_X\_1\_1 / \text{drink}_L\_2\_1
\]

\text{simple-tlink}

As discussed in [Vossen P., 1993], in the multilingual LKB, which is the result of establishing the above tlinks, FSs are linked in terms of formally defined equivalence relations without having checked the equality of their feature-value pairs. As the bilingual dictionary has been used as an external source to get at the tlinks, it may be the case that two FSs are linked as being equivalent which have different feature-value representations.

For example, the source LKB entry chocolate_X_L_2 (which refers to the drinking chocolate), has a number of translations in the bilingual dictionary which refer to both its
liquid and its concrete sense, among other senses. The derived tlink could have been established between the Spanish drinking sense of chocolate and the English concrete sense of the latter, if the classification of the English lexicon was not so fine-grained.

In case a source LKB entry has got more than one translation in the bilingual dictionary belonging to the semantic field of drink, which are mapped to target LKB entries, then all the generated tlinks can be accepted. For example, the two translations that are given for the source LKB entry conac_X_I_1, namely, cognac and brandy are both within the semantic field concerned, and they correspond to the following LKB target entries (i.e. senses): cognac_L_0_0 and brandy_L_1_1, brandy_L_1_2 respectively. Consequently, all the following generated tlinks are acceptable and as such, they are all included in the appendix:

conac_X_I_1 / cognac_L_0_0
simple-tlink

conac_X_I_1 / brandy_L_1_1
simple-tlink

conac_X_I_1 / brandy_L_1_2
simple-tlink

However, when source LKB entries have a big list of possible single translations or translation phrases, in most cases only a small subset of these are part of the target subset. Most of these translations, come from senses in the bilingual dictionaries that do not fall within the semantic field.

The translation given in the bilingual can have various forms; for example, it can be composed, like iced drink for the source LKB entry granizado_X_I_1, or it can be a phrase like sweet milky drink made from chufa nuts or almonds for the source LKB entry horchata_X_I_1, or a phrase with brackets, like claret (wine) for the source LKB entry clarete_X_I_1, or a phrase with a slash, like lime / linden blossom tea for the source LKB entry tila_X_I_3, or it can be a single word in its simplest form, like drink for the LKB source entry bebida_X_I_3.

In the English lexicon, 31 out of the 192 target LKB entries are a concatenation of two or more words, linked by means of an underscore. As we have already explained in section 4.2.3, the existing heuristics can capture such kind of target entries, by means of the compound-dlink-ruleset; based on the translations given in the Harrap’s bilingual dictionary, only one such target lexical entry has been captured, namely, milk_shake_L_0_0 and linked with the source LKB entry batido_X_I_5.
There are cases where words exist in both lexicons, however, they have not been connected by means of a direct link because they are neither included in the bilingual lexicon nor they have exactly the same spelling. For example,

benedictino_X_1_4 / benedictine_L_0_0 or Benedictine_L_0_0
quilanti_X_1_1 / Chianti_L_0_0
pelel_X_1_1 / pale_ale_L_0_0
zarzaparrilla_X_1_2 / sarsaparilla_L_0_0

Sometimes the difference of the source and target LKB entries is as slight as an accent, as shown in the following examples:

anís_X_1_5 / anis_L_0_0
nécantar_X_1_1 / nectar_L_0_1 or nectar_L_0_2

A similar case is shown in the next examples, however, here, the source word exists in the bilingual dictionary and the difference in the accent appears between the translation given in the bilingual dictionary and the target LKB entry:

aperitivo_X_1_2 / aperitif_L_0_0 & aperitif \rightarrow apéritif
(bilingual VOX)
rosado_X_1_3 / rose_L_0_0 & rosado \rightarrow rosé
(bilingual VOX)

Below, another example is given were the mapping from the bilingual dictionary to the target lexicon has failed, due to the inflected form (plural) of the given translation:

bíter_X_1_1 / bitter_L_3_0 & bitter \rightarrow bitters
(bilingual VOX)

The system, at its present state, cannot capture the above cases. Besides, it is still necessary to formally define lexical rules in the LKB (e.g. pluralization) and to construct the corresponding rule sets in the TGE.

Mismatches, in the sense that the corresponding word of a word in the source lexicon is not included in the target lexicon, have also occurred, due to the different taxonomic classification of these words in the two languages. For example,

jarabe_X_1_1 \rightarrow jarabe => syrup \rightarrow NIL (target lexicon)
brebaje_L_1_1 \rightarrow rebaje => beverage \rightarrow NIL (target lexicon)
zumo_X_1_1 \rightarrow zumo => juice \rightarrow NIL (target lexicon)

(the second column shows the bilingual entries and their translations given in the bilingual

Translation Equivalence Via Lexicon: A Study on Tlinks. 22
VOX dictionary)

According to the English taxonomy, syrup_L_1_3 belongs in the taxonomy of oil_L_1_1, beverage_L_1_0 in the taxonomy of liquid_L_2_1, and juice_L_1_2 in the taxonomy of drip_L_2_2. This happens because the genus word that has been selected in their definition ends up in a different to the drink_L_2_1 taxonomy and, obviously, this also affects their taxonomic children.

In most cases, whenever source LKB entries appear two or more times in the appendix, for instance in the list of partial-tlinks generated by means of the parent-tlink-ruleset, this happens because the translation of their parent (via taxonomy) is a polysemous word, corresponding, as such, to two of more target LKB entries (i.e. senses). For example,

galliano_X_L_1 / liquor_L_0_1:
partial-tlink

galliano_X_L_1 / liquor_L_0_2:
partial-tlink

However, in the case of the source LKB entry absenta_X_L_1, this is not the reason of its appearing twice in this list, as shown below:

absenta_X_L_1 / absinth_L_0_1:
partial-tlink

absenta_X_L_1 / absinth_L_0_2:
partial-tlink

but, it is due to the fact that absenta_X_L_1 appears twice in the Spanish taxonomy of bebida_X_L_3, the first time as child of the top bebida_X_L_3 while the second as child of ajenjo_X_L_2.

Another observation on the results, is that for the source LKB entry vodka_X_L_1 which is the taxonomic child of vodka_X_L_1, only partial-tlinks generated by means of the grandparent-tlink-ruleset have been generated by the system (see appendix) and none by the parent-tlink-ruleset because the latter ruleset does not call the orthographic-tlink-ruleset, in the present state of the system.

Further, the next example shows a case of a linkage which would not have been captured with the earlier version of the program PRE:

bebedizo_X_L_3 / potion_L_0_0:
partial-tlink
This partial-tlink has been generated by means of the parent-tlink-ruleset, which has called, in turn, the general-tlink-ruleset.

The total number of tlinks which have been generated and selected for the taxonomy of *bebida_X_L_3 (drink)* with the explained software is 376 tlinks. In particular:

**simple-tlinks**

- generated by means of the simple-tlink-ruleset: 55 (14.5%)
- generated by means of the compound-tlink-ruleset: 41
- generated by means of the orthographic-tlink-ruleset: 13

**phrasal-tlinks**

- generated by means of the phrasal-noun-tlink-ruleset: 2 (0.5%)

**partial-tlinks**

- generated by means of the parent-tlink-ruleset: 320 (85%)
- generated by means of the grandparent-tlink-ruleset: 268
- generated by means of the general-tlink-ruleset: 44

All these derived tlinks for the Spanish taxonomy *bebida_X_L_3* are listed in the appendix. In these lists, multiple links are given when several senses apply (see comments above).

### 2.3 Conclusions.

In the previous section, we have reported and described the results of an experiment of automatically extracting equivalence relations for Spanish and English drink-nouns by using the software developed at the UPC, in Barcelona, namely the TGE. The final process is semi-automatic as the tlink generation is done automatically, while the selection of the desired tlinks is done manually.

All the tlink-rulesets work satisfactorily and, therefore, a considerable part of the subsets has been linked (95% of the source lexicon). However the program PRE tlink-rulesets has only been tested over limited subsets of specific semantic fields. Its potential will be, therefore, proved on a later stage, once its application on larger and less restricted sets of
word senses takes place.
References

[Ageno A. et al., 1993a]

[Ageno A. et al., 1993b]

[Copestake A. & Sanfilippo A., 1993]

[Martí M.A. & Soler C., 1993]

[Samiotou A. (forthcoming)]
Samiotou A. (forthcoming) Performance of cross-linguistic equivalence relations: a lexicon-based approach, Dissertation submitted for the degree of MSc in Machine Translation, University of Manchester

[Sanfilippo A. et al., 1992]

[Vossen P., 1993]
Vossen P. (July 1993) “Extracting Equivalence Relations for a Multilingual Lexical Knowledge Base”, Computer Centrum Letteren, University of Amsterdam, Esprit BRA 7315, ACQUILEX II WP nº 14
LSI-93-1-R "A methodology for semantically enriching interoperable databases", Malú Castellanos.

LSI-93-2-R "Extraction of data dependencies", Malú Castellanos and Fèlix Saitor.

LSI-93-3-R "The use of visibility coherence for radiosity computation", X. Pueyo.

LSI-93-4-R "An integral geometry based method for fast form-factor computation", Mateu Sbert.

LSI-93-5-R "Temporal coherence in progressive radiosity", D. Tost and X. Pueyo.

LSI-93-6-R "Multilevel use of coherence for complex radiosity environments", Josep Vilaplana and Xavier Pueyo.

LSI-93-7-R "A characterization of PF^{NP} \subseteq \text{PP}^{NP[log]}", Antoni Lozano.

LSI-93-8-R "Computing functions with parallel queries to NP", Birgit Jenner and Jacobo Torán.


LSI-93-10-R "Parallel approximation schemes for problems on planar graphs", Josep Díaz, Maria J. Serna, and Jacobo Torán.

LSI-93-11-R "Parallel update and search in skip lists", Joaquim Gabarró, Conrado Martínez, and Xavier Meseguer.

LSI-93-12-R "On the power of Equivalence queries", Ricard Gavaldà.

LSI-93-13-R "On the learnability of output-DFA: a proof and an implementation", Carlos Domingo and David Guajarro.

LSI-93-14-R "A heuristic search approach to reduction of connections for multiple-bus organization", Patricia Ávila.

LSI-93-15-R "Toward a distributed network of intelligent substation alarm processors", Patricia Ávila.

LSI-93-16-R "The Odyssea approach to the design of information systems from deductive conceptual models", Maria Ribera Sancho and Antoni Olivé.

LSI-93-17-R "Constructing face octrees from voxel-based volume representations", Robert Juan i Ariño and Jaume Solé i Bosquet.


LSI-93-21-R “Extending a single resolution system towards a distributed society”, Karmelo Urzelai.


LSI-93-23-R “Especificació d'una biblioteca de tipus”, Xavier Franch (written in Catalan).


LSI-93-26-R “Modelo para el control de calidad en LESD basado en la medición del software”, O. Slávkova (written in Spanish).


LSI-93-29-R “Anàlisi de les definicions verbals del diccionari Vox”, Mariona Taulé Delor (written in Catalan).

LSI-93-30-R “The structure of a logarithmic advice class”, Montserrat Hermo.

LSI-93-31-R “Toward a realistic semantics of possible worlds for logics of belief”, Gustavo Núñez, Matías Alvarado, and Ton Sales.

LSI-93-32-R “Conocimiento en mundos posibles mediante una relación de posibilidad constructiva”, Matías Alvarado (written in Spanish).


LSI-93-34-R “Concatenation versus addition in knapsack problems”, Birgit Jenner.


LSI-93-40-R “How to know it all (or how to get rid of logical omniscience and perfect reasoners)”, Ton Sales.


LSI-93-44-R “The $P_{\log}$ and $AC^{-1}$ operators on the polynomial time hierarchy”, Jorge Castro and Carlos Seara.


LSI-93-49-R “A class-based approach to learn appropriate selectional restrictions from a parsed corpus”, Francesc Ribas.

Internal reports can be ordered from:

Nuria Sánchez
Departament de Llenguatges i Sistemes Informàtics (U.P.C.)
Pau Gargallo 5
08028 Barcelona, Spain
secretsi@lsi.upc.es