HAL
open science

Translation prediction using word cooccurrence graphs

Marianna Apidianaki

To cite this version:

Marianna Apidianaki. Translation prediction using word cooccurrence graphs. Proceedings from The Corpus Linguistics Conference Series, 2005. halshs-00010277

HAL Id: halshs-00010277

https://shs.hal.science/halshs-00010277v1

Submitted on 18 Apr 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Translation prediction using word cooccurrence graphs

Marianna Apidianaki
Lattice
CNRS / ENS / University Paris 7 – Denis Diderot
Marianna.Apidianaki@linguist.jussieu.fr

1. Introduction

Word sense disambiguation (WSD) is a thorny subject in natural language processing. It is implicated in many NLP tasks at varying degrees, where it usually constitutes an intermediate stage of processing and not a goal in itself. Applications relative to translation (machine translation, bilingual lexicon building etc.) are highly concerned with WSD. The polysemy of source (SL) and target language (TL) lexical items influences the strategies adopted during the translation process and the final translation choices. It also complicates the detection of relations between polysemous items and their equivalents in texts, which are rarely one-to-one. The strategies imposed on the translator by SL lexical items are described in Salkie (2002) by a continuum, which goes from those imposed by items that are always translated in the same way in the TL and are thus *translationally systematic*, to those imposed by *translationally asystematic* items, being translated differently every time they occur in texts. These last cases, as well as intermediate ones, are quite demanding for human or automatic treatment and they often require the resolution of lexical polysemy.

Observations coming from bilingual or multilingual translation corpora – consisting of original texts in one language and their translations in one or more other languages – can provide new insights to these questions. Corpus work offers the possibility to empirically test the validity of well-established assumptions about language, providing a better understanding of actual phenomena. The greater availability of corpora of this type makes it possible to extent this kind of analysis in the field of translation as well as in that of contrastive language studies.

In this paper, we will propose a method of translation prediction for polysemous lexical items. The first step in this process will be the disambiguation of SL items, which will subsequently permit the definition of fine-grained translation correspondences. So in cases of multiple translation candidates for new occurrences of polysemous items, the most suitable translation will be found by a combination of monolingual and bilingual information. But let's first take a look at the way polysemy is considered in a translation context.

2. Polysemy in a translation perspective

2.1. The relation between meaning and use in translation

A theoretical framework well-suited for the corpus study of meaning is the contextual approach to meaning, developed by Firth (1957b: 11) in the line of Wittgenstein's conception of meaning (1953: 18). This type of approach marks a shift from the

conceptual approach to meaning and from studies based on introspection to the situational approach and the consideration of meaning in use[1]. This change of orientation in linguistic semantics came about almost at the same time as the change of the status of meaning in translation studies and the decline of the semantic view of the relationship between source and target texts. The idea that meaning consists of something stable and pre-existing that can be recovered from texts in one language and transferred in texts of another language *in the same way as one might transfer wine from one glass to another* is abandoned, as well as the assumption of separation of form and meaning (Baker, 1993: 236). So, in the framework of translation, Firth (1968: 91) suggests connecting structures and systems of languages to structures and systems in situations in which language functions and Haas (1968: 104) considers correspondence in meaning as correspondence in use. The more intense preoccupation with meaning in use, at a less abstract (conceptual) level, is also reflected in the passage from interlingual to transfer methods in the field of Machine Translation (Fuchs *et al*., 1993: 206-209).

The notion of translation equivalence is reassessed too; it no longer refers to a static relationship between source and target texts. The search of stylistic and functional equivalence undermines the primacy of the original text over the translation and gives the TL texts a new status in translation studies. This change in orientation imposes the study of a large number of authentic texts in both languages and prepares thus the ground for descriptive and corpus work on these subjects.

2.2. Overlapping vs. diverging polysemy

Exploring polysemy in a bilingual context is a complicated but rather informative task. Languages divide up semantic space in different ways, conceptual structures evolve differently and complete equivalence is rare. This picture induces various cross-linguistic relationships, such as those of *overlapping* and *diverging polysemy* (Altenberg and Granger, 2002). In the first case, items in two languages have roughly the same meaning extensions, while in the case of *diverging polysemy* items' meaning extensions vary.

A consequence of diverging (or partially overlapping) polysemy is that words treated as translation equivalents in dictionaries often have different meaning extensions or ranges of meaning. These items display thus low mutual correspondence in texts due to their divergent meaning extensions (Salkie, 1997; Viberg, 2002), which result in a wide range of translations. Another type of cross-linguistic relation is that of *no correspondence* at the semantic or conceptual level, which renders the task of finding obvious translation equivalents very difficult. Lack of correspondence usually results either in zero translations or in multiple translations highly dependent on context.

Given the complexity of the situation, we assume that any application relative to translation should take into account the intricate relations existing between languages at the level of semantics if it is supposed to give satisfactory solutions. For this to be done, relations between lexical items should be established at a *lower* level than that

[1] Here, we'll limit our survey to the contribution *collocation* can have to the analysis of meaning, leaving the other aspects of contextual meaning out of the scope of this work.

of words. Such relations could describe more clearly the correspondences existing between polysemous items. Nevertheless, with this analysis we do not intend to reduce cross-linguistic equivalence to a matter of semantic content but to explore just one aspect of it.

2.3. Polysemy and translational ambiguity

More often than not, the study of polysemous SL items and of their multiple translation equivalents provides valuable linguistic and translation information. It has been demonstrated (Gale *et al.*, 1992, 1993; Dagan *et al.*, 1991; Teubert, 2002) that it can serve the disambiguation or, even, the sense annotation task for polysemous SL items. However, the authors do not forget to mention the limits of this approach, in cases where word-sense ambiguity is preserved across languages, i.e. the TL equivalents are ambiguous too. These are cases where the ambiguous SL items are not *translationally ambiguous* (Salkie, 2002). The occurrence of such cases depends highly on the target language; they are observed most often in closely related languages, where items are multiply ambiguous in more or less the same ways (ex. En: *interest*, Fr: *interêt*). In such cases a third language is needed for disambiguation.

On the other hand, the existence of multiple translation equivalents is not necessarily indicative of a sense split in the SL. Given that languages divide semantic space differently, it may happen that distinctions appearing in one language do not exist in the other and that a more generic term in one language is used to express more than one sense, which in the other language may be expressed by various lexical items. But this is not the only case where the use of multiple translation equivalents does not reflect sense distinctions in the SL. Translators sometimes exchange the use of synonyms or near-synonyms for stylistic reasons, quite often in order to avoid repetitions in the TL text or for originality.

The distinction of cases where the use of multiple translation equivalents is due to the polysemy of source lexical items from those where it is due to other factors, such as stylistics, is not always evident. However, linguistic evidence coming from SL and TL texts could possibly shed some light on this question. Regularities of use in the SL coupled with more or less systematic and regular use of the multiple TL equivalents could imply the existence of distinct senses in the SL. Testing such a hypothesis implies a combination of monolingual and bilingual information and requires a large corpus providing a multitude of examples. On the other hand, in cases where stylistics is involved, we would expect translation choices to be more random, less regulated by linguistic evidence and mostly by factors such as the idiosyncrasy of translator, his talent and his personal taste. Nevertheless, even in this case some regularity may appear within a given corpus from a single translator. The proposed method accounts for both cases and takes appropriate actions to deal with them.

In order to explore our premises concerning the translation of polysemous items, we first have to look for their occurrences and their translation equivalents in texts, which often is a 1: *n* relation. These equivalents may play a more or less important disambiguating role *vis-à-vis* the corresponding SL items, depending on their own degree of polysemy. The investigation of the relations between source and target items in actual translations and of those holding in both parts of the corpus between

them and the lexical items surrounding them in texts can provide useful clues for the disambiguation procedure.

2.4. Automatic WSD for translation prediction

Automatic WSD can have useful applications in a translation context. In a Machine Translation framework, it can be of great help for the choice of the most adequate translation in cases of multiple possible equivalents. Another application could be a translation prediction module integrated in Computer-Assisted Translation (CAT) tools. It is however true that human translators are not always confronted with multiple translation candidates in cases of polysemous SL items. In some cases, the appropriate meaning of words is effortlessly selected and the context "imposes" one of the equivalents in such a way that the other possibilities do not even cross the translator's mind. However this is not always the case. Automatic WSD would permit to the system to make translation suggestions taking into consideration the linguistic context of words. For this, monolingual and translation information on past and new occurrences of lexical items would be needed.

The input for the translation prediction process could be the result of a word alignment or translation spotting task, or the entries of a bilingual dictionary (or glossary, or terminological database). These would serve as translation candidates. The next step would consist in finding the most suitable translation in context combining bilingual with co-textual information. The translator would then have to decide if he would incorporate it or not in his translation. Such a tool would thus quicken and facilitate his work.

In this paper we will explore the possibility of using an existing automatic WSD method for disambiguation in a translation context. This method was presented in Véronis (2004) and was used in a monolingual information retrieval perspective. In our work, the results of the disambiguation procedure on the monolingual side will be refined using bilingual information. Then the relations existing between polysemous items in both languages will be explored by means of an example of a polysemous word having various renderings in the TL. This study does not claim to be exhaustive. The method proposed presents some advantages for the treatment of polysemy in a bilingual context, casting light on the relations that can exist between linguistic evidence from actual texts in the SL and multiple translation equivalents in the TL. So, it permits the establishment of finer-grained correspondences between lexical items than correspondences at the word level.

3. Cooccurrence graphs for sense distinction and disambiguation

3.1. Turning cooccurrence information into graphs

In the framework of the contextual approach to meaning, cooccurrence information can serve for the disambiguation of polysemous lexical items. So the various senses of polysemous words can be represented by means of a cooccurrence graph (Véronis, 2004), where nodes represent words and links are perceived as significative

cooccurrences between words in texts. The edges connecting two words are weighted and those having a weight >0.9 – which means that the words are not strongly related – are eliminated. The weight of the edges is calculated by the formula

$$w_{A,B} = 1 - \max\,[p(A|B), p(B|A)]$$

where p(A|B) is the conditional probability of observing A in a context that contains B and the inverse for p(B|A). These graphs are of type *small world* and, because of their structure, they lie somewhere between random and regular graphs (Watts & Strogatz, 1998). A characteristic of this kind of graphs is that most nodes have few connections, while a small number of nodes – the root hubs – are highly connected to a large number of others. Small world networks depict an important property of human language, i.e. the fact that any word in the lexicon can be reached with fewer than three intermediate words on average (Ferrer *et al.*, 2001).

In a cooccurrence graph, *high density components* – areas where there are many connections between the nodes – represent the various senses of lexical units. Each component has a *root hub*, which is the node with the highest degree in the component. The hubs help to delineate the high density components, i.e. the different senses. It should be noted that this method – as presented in (Véronis, 2004) – is well suited for disambiguating content words and, especially, nouns and adjectives. Verbs deteriorate the results of the disambiguation task because their uses are often too general and cannot serve as strong disambiguators for other lexical units. It would however be very interesting to explore the contribution of content words of other grammatical categories in disambiguation.

3.2. Building the cooccurrence graph – an example

The elements used as "bricks" for the construction of the graph are the cooccurrences (nouns and adjectives) of the polysemous SL item – here, the noun *plant*. The graph has been created manually using the instructions given in Véronis (2004) with some deviations, such as the thresholds adopted. The occurrences of the noun *plant* have been extracted from the INTERA corpus (Gavrilidou *et al*., 2004). This is a 4.000.000 words corpus, which contains English texts coming from five different domains (Education, Health, Law, Environment and Tourism) and their translations into Greek. All texts are taken from the Official Journal of the European Union. Despite the large size of the corpus the lexical density is quite low, given that Community texts are characterized by a high degree of repetition.

The graph was created using the frequency lists of *plant*'s cooccurrences – and of their respective cooccurrences – as well as the list of the total frequencies of the lexical items in the corpus. The components were constructed taking into account even words cooccurring twice with the target word. This threshold depends highly on the type and size of the corpus.

The graph offers the possibility to visualize the relations between words describing the various senses. Grouping *plant*'s uses, we initially discern two senses. In terms of cooccurrence information, the word is used with one sense when it cooccurs with the items {*variety, species, seed, catalogue, shrub,…*} and with another one when it

cooccurs with some of the following: {*power, nuclear, reactor, Ignalina, decommissioning, emission, combustion, dioxide,…*}. The order of description of the senses corresponds to the frequency of appearance of the corresponding uses in the corpus. We do not exclude – or, better, we consider highly probable – a different ordering of these senses or the detection of different ones, in the case of a corpus of another type or of different size. In Figure 1 we describe the component that corresponds to the second sense of *plant*:

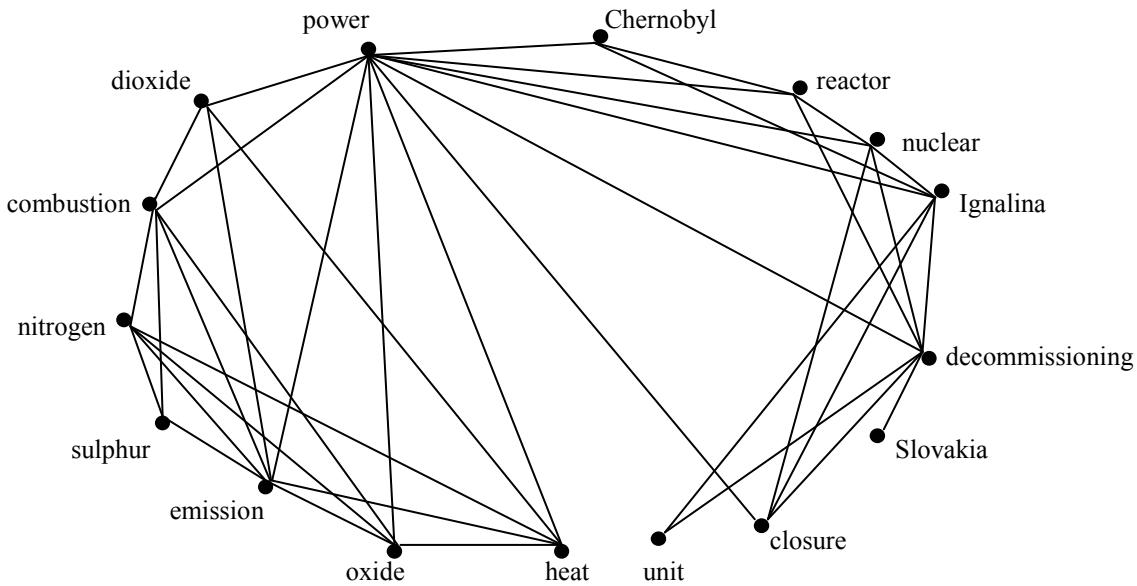

Figure 1. A high density component of the graph of *plant*

The component's root hub is *power*. It is the node with the highest degree and the highest frequency. For the disambiguation of a new occurrence of the polysemous / target word, it is not necessary that it cooccurs with *all* the elements described in the graph; even one occurrence of a good disambiguator – i.e. a word that has a close relation to the target word and that appears in the graph – may suffice. Nevertheless, it is quite common to find occurrences of the target word whose cooccurrences are not included in the graph. In our example, this happens with words that appear frequently in the corpus (like *way, value, animal, health, protection,* as well as words relative to the structure and activities of the European Union, like *member, state, article, directive, treaty…*), so the thresholds impede them from entering the graph.

3.3. Building the Minimum Spanning Tree (MST)

The next stage of the procedure consists in the creation of the MST corresponding to the components of the graph (Figure 2). The polysemous target word – here, the word *plant* – is the root of the tree and the hubs of the components constitute its first level. The branches of the tree correspond to the components previously detected and they represent the two senses. An important characteristic of the MSTs from a linguistic point of view is that they point out the kind of relations existing between lexical items. These can be primary relations (*seed, catalogue*) or secondary ones (*hybrid,*

tropics) occurring by means of transitivity, following the small world principle "*The friends of my friends also become my friends*" (Véronis, 2004). The exploitation of this information could prove very useful in the translation process through an appropriate weighting schema.

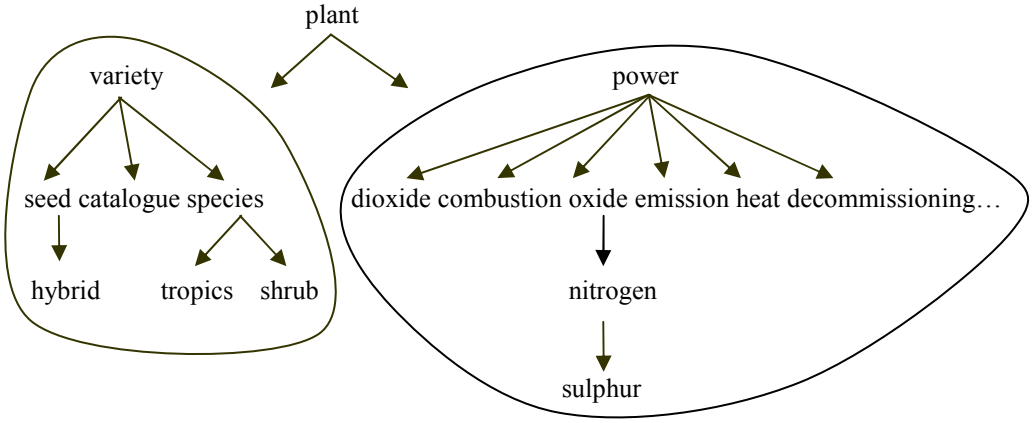

Figure 2. The Minimum Spanning Tree of *plant*

We have to note here that if a graph is constructed for each language (SL and TL), a network of links could be created, which would allow the modeling of sense-to-sense correspondences describing the relation of equivalence between source and target items – *sense* as used in the framework of the contextual approach to meaning. This is very important in cases of multiple translation correspondences between lexical items, where word-to-word correspondences are coarse-grained and often ambiguous or vague. In the following sections we will see how we can give a bilingual dimension to the SL graph without yet constructing a graph for the TL. Taking into consideration the TL information the SL graph can be refined, allowing the TL to play a complementary role of disambiguator for distinctions that are not evident in the SL.

4. Spotting translation equivalents

As we have already seen, the input for the translation prediction process could be the result of translation spotting. In cases of multiple possible equivalents, these would serve as translation candidates. Translation spotting, as defined in Véronis and Langlais (2000) and Simard (2003), is the task of identifying the words in a TL text that correspond to some given words in the SL text. It is considered as a sub-problem of word alignment, as the objective is not to align all words in texts but only a subset of them. The input to this process is a pair of SL and TL text segments and a query – the set of words we wish to align – and the output is a set of tokens in the two languages.

However, translation spotting for words having multiple acceptable equivalents is quite complicated. Frequency of usage of the equivalents may vary greatly, which renders the task of choosing more than one acceptable equivalent quite difficult for statistical alignment tools. Searching for a second (or third etc.) correct equivalent may leave in too much noise, except for rare cases of equivalents being used with similar frequency. Another parameter that complicates this task is that some of the

equivalents appear quite rarely, so they are often "buried" under the frequency thresholds of translation spotting tools.

Here we have manually found the translation equivalents of *plant* in the corpus. First, we had to determine the size of the text segments in which they would be searched. We chose to look for them in translation segments at the sentence level; this permits us to link the shift of meaning to evidence in the near co-text of words rather than in whole texts, which is preferable given that the meaning of words can change even in the same text. Sentence correspondences can be 1:1, 1:2, 2:1 or 2:2 (in order to capture crossing correspondences). These are the types of correspondences permitted by most sentence alignment tools. The correspondences between *plant* and its TL equivalents found in the corpus are illustrated in Figure 3.

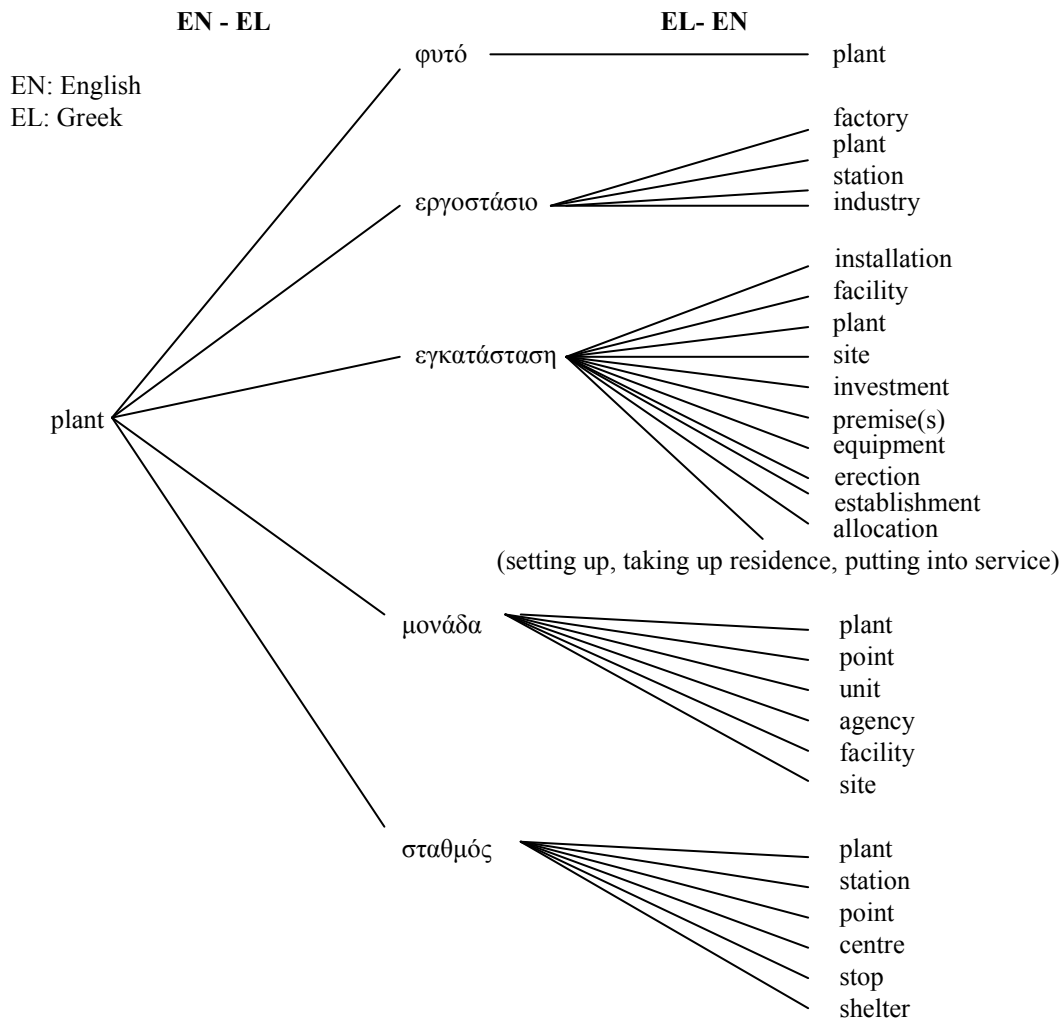

Figure 3. A case of diverging polysemy

In the previous section we described the two senses of *plant* that were revealed by its uses in the English side of the corpus. The translation equivalents found in the Greek texts are five, which means that different translations do not necessarily correspond to different senses. We can thus separate the TL equivalents that denote a sense split in the SL from those that are not. We should also note that there is a great overlap between the domains regarding the use of the various translation equivalents.

Another point to make is that some of the translation equivalents of *plant* are polysemous in the target language. So we are confronted with a classical case of *diverging polysemy*. Reversing the direction of translation spotting, we find multiple translation equivalents for the Greek items in the corpus (Figure 3), which may also denote different senses or not. For the moment we will limit our analysis to the English noun *plant*, assuming that the same kind of treatment could be adopted when reversing the translation spotting direction.

5. Cooccurrences of source and target language items

The translation equivalents of the noun *plant* in Greek that appear in the corpus are: *φυτό, εργοστάσιο, εγκατάσταση, μονάδα, σταθμός* (Figure 3). In Table 1 we can see the frequencies with which *plant* is translated by its various Greek equivalents, as well as some of its most frequent English cooccurrences in each case. In total, it appears 181 times in the texts and its most frequent translation is *φυτό*. Some of its cooccurrences when translated as *φυτό* are included in the component of the graph formed by the words {*variety, seed, catalogue, species, hybrid, tropic, shrub*}. There are some cases where *plant* has these cooccurrences but is translated as *φυτικός*, which is not a noun but an adjective. The use of equivalents of a different grammatical category is quite common in translation. One way to handle this in this context is to consider the adjectival equivalent as an alternative translation of *plant* when used in this particular sense and leave to the translator the work of choosing the right one using more information from the co-text.

	Greek equivalents	**English cooccurrences**
1.	φυτό (38,12%)	{animal, variety, health, species, protection, human, seed, land, medicinal, water, use}
2.	σταθμός (17,12%)	{power, nuclear, Ignalina, unit, decommissioning, Lithuania, measure, closure}
3.	μονάδα (15,46%)	{emission, creation, company, dust, part, way, value, limit, nitrogen, oxide}
4.	εγκατάσταση (11,60%)	{emission, value, combustion, limit, derogation, sulphur, oxide, dioxide, nitrogen, heat}
5.	φυτικός (9,94%)	{variety, species, agricultural, product, catalogue, planet, spread, harmful}
6.	εργοστάσιο (7,73%)	{power, nuclear, Chernobyl, environmental, waste, water, environment}

Table 1. Frequency of use of the TL equivalents and corresponding English cooccurrences of the noun *plant*

When *plant* cooccurs with the words included in the component described in Figure 1, it can have four different translation equivalents in Greek: *σταθμός, μονάδα, εγκατάσταση* and *εργοστάσιο*. We observe that although *plant* gets four different TL equivalents, its SL cooccurrences in all cases are found in the same high density component of the graph, which means that they describe the same sense. So, an assumption to be tested is whether in cases like this the meaning of the TL equivalents is similar to that of the SL item and are thus synonyms or quasi-synonyms in the TL.

If this is the case, we can assume that the translator alternates their use for stylistic reasons.

An alternative is to go a bit further in the distinction of senses exploiting translation information. Looking at the elements that cooccur with *plant* when it is translated as *σταθμός* or *εργοστάσιο* we can see that the intersection of the two sets is not empty, i.e. there are elements in common between the two sets. Similarly, the intersection of the sets of *plant*'s cooccurrences when rendered as *μονάδα* or *εγκατάσταση* is not empty either. On the contrary, the intersection of the two derived sets of elements is empty, i.e. there are no common elements. We interpret this as the possibility of existence of another sense distinction reflected in the use of the different equivalents but not present in the graph.

Turning to the graph now, we can see that the structure of the largest high density component (Figure 1) is such that we could easily distinguish two smaller components in it. These components have many interior connections but are not highly connected between them, as their only common node is *power*. In order to separate them, we keep the node *power* as the root hub of the component in which it has the highest degree and we consider as root of the other component the node with the highest degree in it. We could thus envisage using, on the one hand, translation information and, on the other hand, information coming from the structure of the graph to proceed to the distinction of "finer" senses in the SL. If we do that, the MST of *plant* will change (Figure 4).

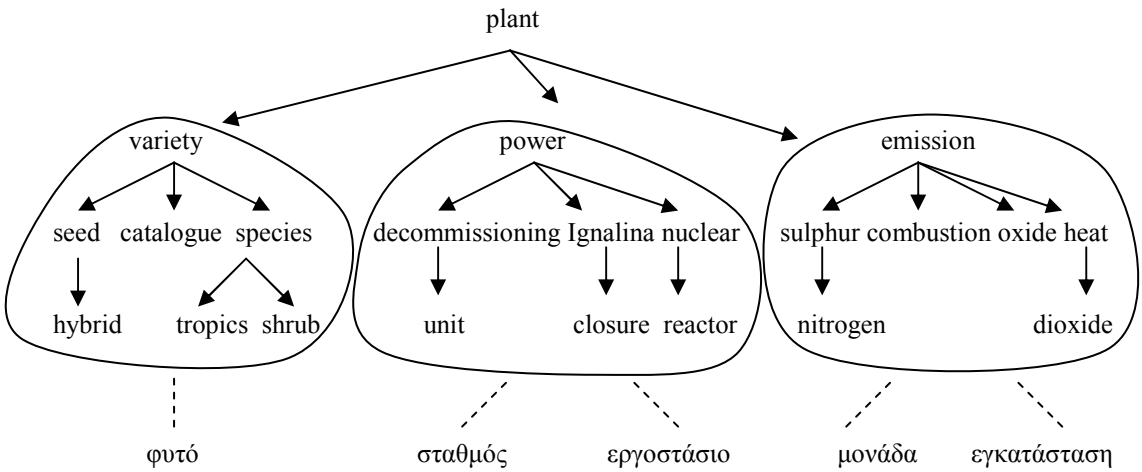

Figure 4. The MST of *plant* reflecting sense distinctions induced by TL items and the corresponding translations in Greek

We looked this sense distinction up in the COLLINS COBUILD English dictionary and in the multilingual term bank of the EU Commission's Translation Service EURODICAUTOM. In the monolingual dictionary we find the distinction between "*a factory or a place where power is generated*" and "*large machinery that is used in industrial processes*". In EURODICAUTOM, *plant* gets the translation *εργοστάσιο* when it refers to "*an industrial facility where raw materials or semi-manufactured products are turned into end-products*". On the other hand, it is translated as *εγκατάσταση* when used in the sense "*the land, buildings, machinery, apparatus, and*

fixtures employed in carrying on a trade or a mechanical or other industrial business" or "*any establishment or other stationary plant used for industrial or public utility purposes which is likely to cause air pollution*".

Finally, we would like to come back to cases where cooccurrences of the polysemous target word in texts are not found in the graph. This often happens with words that are used very frequently in the corpus, so they are not related to the target word strongly enough to be included in the graph (their edges get a weight above 0.9). This is a weak point of this approach: in cases where the target word cooccurs with general or very frequent words, it gets very difficult to identify its sense. In those cases we could possibly try to enlarge our window on texts and take into account the previous or next translation segment, in order to find disambiguation clues that would help to identify the correct sense.

6. Translation prediction

The combination of monolingual and bilingual information coming from existing texts could form the basis for automatic translation prediction. The integration of such a module in a CAT tool can be of help to the translator either by making him/her unique translation suggestions for polysemous SL items, or by restricting the number of possible equivalents and facilitating the final choice.

The first step of this process would be the disambiguation of the new occurrence of the polysemous target word to be translated. This could be done using information from the linguistic context of the word in the new text and information in the MSTs, following the disambiguation procedure described in Véronis (2004). Once the sense of the target word detected, the next step would consist in the search of the most adequate translation equivalent for this sense. If a strong relation exists between the detected sense of the target word and a particular translation equivalent, this one could be suggested to the translator.

In cases where multiple translation equivalents correspond to one sense described in the graph, we assume that the choice of one of them is regulated by stylistic reasons. This means that we cannot find sufficient linguistic evidence in the texts justifying the use of one of those equivalents. So, we could envisage suggesting all of them in cases where their use does not reflect or does not induce a sense split in the SL, leaving to the translator the choice of which one to use on the basis of his stylistic preferences.

7. Conclusion and perspectives for future work

The exploration of polysemy in a bilingual context is quite a complicated task, given the divergent ranges of meaning of words treated as translation equivalents. Mutual correspondence of assumed translation equivalents having different meaning extensions is rather low and it strongly depends on the linguistic context surrounding them in texts. In this paper we have proposed a method that could give the possibility of establishing finer-grained translation correspondences between lexical items than word-to-word ones, which leave in a great deal of ambiguity and vagueness. These

finer-grained correspondences are assumed to reflect sense-to-sense relations. Different senses of SL items are represented in a graph whose subgraphs represent the various senses and are put into correspondence with the different translation equivalents.

Although this method presents some advantages for the treatment of polysemy in a translation perspective, there are many pending issues which would merit further exploration. One such issue is the creation of a TL graph that would allow better description of sense-to-sense correspondences, taking into account the polysemy of the TL and its linking to the SL graph. The creation of a TL graph would also allow the reversion of the translation prediction direction, from the TL to the SL.

Another question to be investigated concerns the thresholds to be adopted for the construction of the graphs. These are highly dependent on the type and size of the corpus and it would be interesting to explore the influence of various thresholds on disambiguation and translation prediction.

The degree to which information from the TL should intervene with sense distinctions in the SL is also an issue to be explored. There are cases where the use of multiple equivalents clearly denotes sense distinctions in the SL and others where it is quite easy to conclude that we have to do with lexical variation due to other reasons. However, it is not always easy to draw the line where we should stop looking in the TL for evidence supporting sense distinctions that are not directly evident in the SL.

We could make another remark regarding the relations of TL equivalents to SL components. We should expect to find cases where a TL equivalent corresponds to more than one component of the SL graph. The investigation of the way of capturing and describing this relation will constitute part of future work. We could, for example, envisage using a metric in order to measure the relation or distance between a TL equivalent and the corresponding SL components and then establishing weighted links between them, which would show more clearly their association strength. This could assist the translation prediction process.

Cases of complex terms are also worth taking into consideration. These terms could be identified in advance, before the construction of the graph. They would then be used in the graph as wholes and contribute to the disambiguation of other lexical items. Most often complex terms would not have to be disambiguated as their polysemy is limited and their different translations are usually term variants in the TL.

Alternatively, we could consider the possibility of identifying complex terms exploiting the information that is inherent in the graph. Nodes are linked by weighted edges, so we could expect to find a high correlation between components of complex terms. Apart from that, translation information can be extremely useful for determining such a relation between SL items. However, for this to be done we have to determine very carefully the thresholds that will be used, in order to ensure the inclusion in the graph of elements that constitute components of complex terms.

Looking at the example described in this paper, we find in the corpus occurrences of the sequences *plant protection, plant growth* and *cutting plant* which are respectively translated in Greek by the compounds *φυτοπροστασία* and *φυτογεωγραφία* and by the

complex term *εργαστήριο τεμαχισμού*. Combining information on the cooccurrence of the English lexical items with information on their translation, we assume that we could more confidently identify these sequences as having a terminological status as wholes[2]. A major advantage of such an approach is that it allows the TL to play a role on the identification of translation units. So, without making any preliminary assumptions on the nature and length of these units we let the data guide us in this respect.

Another question that poses itself is what has to be done when cooccurrence information in new text segments is not enough or appropriate to proceed to disambiguation. We suggested a solution which consists in enlarging the text window, leaving in text from neighbouring translation segments. How far can we go with that and at which moment further action to ameliorate the disambiguation method is required are issues to be explored.

The proposed method could be implemented both dynamically and statically. Dynamic implementation would be more appropriate in the case of continuously and frequently updated translation corpora and would also allow the user to perform his own queries thus being more adaptable to his needs. The obvious disadvantage of such an implementation would be its cost regarding necessary computational resources and response time. If we envisage the a-priori creation of a bilingual resource based on existing bilingual corpora, we loose the interactive character of the dynamic approach. In this case, correspondences found will be pre-tailored to available corpora and less adaptable to users' specific needs. Another issue that should be considered in this case is the required effort of keeping up-to-date the bilingual resource. The choice of one of the alternatives is a matter of the envisaged application and of the importance given to the role of the user and to the regular updating of the data. The evaluation of the performance of the method in each case will constitute part of future work.

[2] We should however note that the words *protection, growth* and *cutting* are not included in the graph of *plant* because of the threshold of association adopted.

References

Altenberg, B. and Granger, S. (2002) Recent trends in cross-linguistic lexical studies, in B. Altenberg and S. Granger (eds.) *Lexis in Contrast, Corpus-based approaches* (Amsterdam / Philadelphia: John Benjamins Publishing Company), 3-48.

Baker, M. (1993) Corpus linguistics and translation studies, in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology, In Honour of John Sinclair* (Philadelphia / Amsterdam: John Benjamins Publishing Company), 233-250.

Dagan, I., Itai, A. and Schwall, U. (1991) Two Languages Are More Informative Than One. *29th Annual Meeting of the Association for Computational Linguistics (ACL), Berkeley, California, 1991*, 130-137.

Ferrer I Cancho, R., Solé, R. V. (2001) The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, *268(1482)*, 2261-2265.

Firth, J. R. (1957a) *Papers in Linguistics, 1934-1951* (London / New York: Oxford University Press).

Firth, J. R. (1957b) A Synopsis of Linguistic Theory, 1930-1955, in *Studies in Linguistic Analysis,* Special Volume of the Philological Society (Oxford: Basil Blackwell).

Firth, J. R. (1968) Linguistics and Translation, in F. R. Palmer, *Selected Papers of J.R. Firth: 1952-59* (London: Harlow: Longmans), 84-95.

Fuchs, C., Danlos, L., Lacheret-Dujour, A., Luzzati, D., Victorri, B. (1993) *Linguistique et traitements automatiques des langues* (Paris: Hachette).

Gale, W. A., Church, K. W. and Yarowsky, D. (1992) Using Bilingual Materials to Develop Word Sense Disambiguation Methods. *Fourth International Conference on theoretical and methodological issues in machine translation*, *Montreal, 1992*, 101-112.

Gale, W. A., Church, K. W. and Yarowsky, D. (1993) A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities 26(5)*, 415-439.

Gavrilidou, M., Labropoulou, P., Desipri, E., Giouli, V., Antonopoulos, V. and Piperidis, S. (2004) Building parallel corpora for eContent professionals. *MLR 2004, PostCOLING Workshop on Multilingual Linguistic Resources, Geneva, 2004*.

Haas, W. (1968) The theory of translation, in G. H. R. Parkinson (ed.) *The theory of meaning* (London: Oxford University Press), 86-108.

Kenny, D. (2001) *Lexis and Creativity in Translation: A corpus-based study* (Manchester: St Jerome Publishing).

Salkie, R. (1997) Naturalness and contrastive linguistics, in B. Lewandowska-Tomaszczyk and P. J. Melia (eds.) *Proceedings of PALC 97: Practical Applications in Language Corpora* (Lodz: Lodz University Press), 297-312.

Salkie, R. (2000) Quelques questions méthodologiques dans l'exploitation des corpus multilingues, in M. Bilger (ed.) *Corpus: méthodologie et applications linguistiques* (Paris : Honoré Champion), 180-195.

Salkie, R. (2002) Two types of translation equivalence, in B. Altenberg and S. Granger (eds.) *Lexis in contrast: Corpus-based* (Amsterdam / Philadelphia: John Benjamins Publishing Company), 51-71.

Simard, M. (2003) Translation Spotting for Translation Memories. *NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, Canada, 2003*.

Teubert, W. (2002) The role of parallel corpora in translation and multilingual lexicography, in B. Altenberg and S. Granger (eds.) *Lexis in contrast: Corpus-based approaches* (Amsterdam / Philadelphia: John Benjamins Publishing Company), 189-214.

Tognini-Bonelli, E. (1996) Towards Translation Equivalence from a Corpus Linguistics Perspective. *International Journal of Lexicography, 9(3),* 197-217.

Véronis, J. and Langlais, P. (2000) Evaluation of parallel text alignment systems – The ARCADE project, in J. Véronis (ed.) *Parallel Text Processing* (Dordrecht: Kluwer Academic Publishers), 369-388.

Véronis, J. (2004) Hyperlex: lexical cartography for information retrieval. *Computer, speech and dialogue, 18(3)*, 223-252.

Viberg, Å. (2002) Polysemy and disambiguation cues across languages: the case of Swedish *få* and English *get*, in B. Altenberg and S. Granger (eds.) *Lexis in Contrast, Corpus-based approaches*, (Amsterdam / Philadelphia: John Benjamins Publishing Company), 191-150.

Watts, J. W. and Strogatz, S. H. (1998) Collective dynamics of small world networks. *Nature 393*, 440-442.

Wittgenstein, L. (1953) *Philosophical Investigations*. (Oxford: Blackwell Publishing).