JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Cornelia Wermut, T. (2008). Linguistics-based word alignment for medical translators. *The Journal of Specialised Translation, 9*, 20-38. https://doi.org/10.26034/cm.jostrans.2008.682

Linguistics-based word alignment for medical translators

Tom Vanallemeersch & Cornelia Wermuth
Lessius University College, Antwerp

ABSTRACT

Tools assisting professional translators memorise translated sentences but provide limited functionality for the identification of terms and their translation equivalents in translated texts. In this paper, we propose a word alignment approach aiming to improve efficiency and usability of this functionality, through identification of cognates and exploitation of linguistic knowledge, such as lemmas in bilingual glossaries and dictionaries. Our approach focuses on content words, is applicable to parallel texts of various sizes, and minimises the need for user parameter tuning and preprocessing steps. The method, implemented as the FragmALex system, tackles certain types of divergences between source and target text by creating and grouping links between *fragments* (word parts, words and word groups). The system output consists of fragment links in their original context. We performed a case study of Dutch and French medical articles, using a medical glossary and a general-purpose dictionary of restricted size.
Comparison of the output with a gold standard shows that the addition of the dictionary to the system accounts for a higher increase in recall (completeness of alignment) than the addition of the glossary, while the decrease in precision remains low with either resource.

KEYWORDS

Translation, parallel texts, sentence alignment, word alignment, terminology extraction, lexical knowledge, medical Dutch, medical French

1. Introduction

Professional translation practice features numerous digital resources, such as bilingual glossaries and previously translated material, allowing the translator to work more efficiently and consistently. Commonly available systems on the market provide automated means of exploiting these resources, but they do so in a rather coarse way. A major drawback is that they store sentences and their translations as strings, without detecting term correspondences between them. The input of bilingual terms is carried out manually in order to allow the software to recognise source terms and their translations in new sentences. These equivalents can also be identified through concordances in existing translated material. The key to efficient detection of term correspondences is automated word alignment. Word alignment is mostly performed using a statistical approach. However, aspects such as the importance of text size and the tuning of parameter values make this approach inappropriate for translators. We therefore propose a method based on linguistic resources such as bilingual dictionaries and glossaries, focusing on efficiency and usability. We apply this method to the field of medical translation, starting from French-Dutch parallel texts.

This paper is structured as follows. In section 2, we describe the typical professional translator's environment and some approaches to sentence

and word alignment. In sections 3 and 4, we discuss our linguistics-based method. Section 5 describes our case study for medical texts, section 6 its results. In the last section, we present our conclusions and future research.

2. Background

2.1. Translator's environment

Professional translators, especially those working in specialised translation, make intensive use of digital resources. These include lexical data (online dictionaries, glossaries and term databases) and translation memories. Another very important resource is the Internet, as static resources like dictionaries only cope to a certain degree with the rapid development of specialised terminology.

Computer Aided Translation tools (Bowker 2002), commonly referred to as CAT tools, automatically store sentences and their translations in a Translation Memory (TM) during the translation process and allow for reusing them in future translation projects or in the same project. This increases translation efficiency and consistency, as the translation of a sentence already present in the TM requires solely revision, and the translation of a similar sentence requires solely adaptation. CAT tools also increase translation consistency by allowing the user to enter terms and their translation into a term database. This process either takes place as a preparatory step before translation (creation or import of a term collection representative of the domain) or incrementally (during the translation process). Such a database allows for term recognition: as soon as the user starts translation, the software checks whether the sentence contains terminology from the database and, if so, suggests its translation.

In spite of their obvious usefulness, CAT tools cannot automatically detect terms and their translations in the TM, such that term correspondences must be entered manually into the term database. The distinction between TM and term database is confusing for first-time users of CAT tools. Translators view the input of terms into the database, even when keyboard shortcuts are available, as an additional workload, and therefore do not perform this process in a systematic way. An alternative method to find out a term's translation is the concordance functionality, which identifies source sentences in the TM containing a specific word or word group, and lists these source sentences together with their translations. Though very helpful, this functionality is time-consuming since the user needs to read through the translations of the sentences in order to find the translation equivalent of the word or word group.
The need for efficiently compiling mono- or bilingual terminology lists from parallel text data such as translation memories has given rise to programs for semi-automatic term extraction. These modules are stand-alone or part of a CAT tool suite. The term extraction functionality is applied as a

preparatory step before the translation process. Most modules perform bilingual term extraction on a statistical basis, restricting their usefulness to high-frequency terms and leading to linguistically unmotivated terms. Other modules are exploiting information on *chunks*; for a discussion of this type of unit, see Planas (2000).

The term extraction tools discussed above implement a specific type of word alignment if applied to bilingual texts. Word alignment will be further discussed in section 2.3.

2.2. Types of divergences

Source texts and their translations display a certain degree of parallelism. Divergences between them arise during the translation process because of linguistic issues, text genre and personal choices of the translator. Compositions in one language (e.g. in Dutch) may be translated by a sequence of words in another (e.g. English, French). The order of words or word parts in one language may be different from that in the target language, e.g. Dutch *muziekwinkel* and English *music shop* vs. French *magasin de musique*. Linguistic divergences and translators' choices are also reflected in other ways. So, the translator may use syntactic categories different from the source text. In example 1, a Dutch verbal construction translates a French nominalisation [1]. Even if words have the same syntactic category in the source and target text, they may not display the same grammatical features. For instance, an order expressed by an infinitive may be translated by an imperative.

(1) FRA Pendant l'*emballage* des produits
NLD Terwijl we produkten *verpakken*
ENG 'While we pack products'

In addition, some languages clearly differ as to the 'wording' (e.g. French tends to be more verbose than Dutch).

Text divergences between larger text areas may consist of reordered, added, omitted or paraphrased sentences or paragraphs, due to layout issues or personal translator's choices. For instance, if a source sentence starts with a subordinate sentence, the latter may appear at the end of the corresponding target sentence. One sentence may be translated by several ones or vice versa.

2.3. Alignment

The principle of alignment is the automated identification of correspondences between elements of the source text and its translation, like paragraphs, sentences and words. This is a non-trivial task, due to the cross-lingual divergences described in the previous section.

Research focusing on sentence alignment (Gale and Church 1991, Kay and Röscheisen 1993) showed that it is feasible to tackle common phenomena like a translation of one sentence by multiple ones. From a practical point of view, layout markup provides useful information; see Ricca, Tonella, Pianta and Girardi (2004). Word alignment, especially in case of poor parallelism, proves to be rather difficult. Generally, it is applied after sentence alignment, and statistically approached, for instance through statistical machine translation (Brown, Della Pietra, Della Pietra and Mercer 1993, Och and Ney 2000b).

Word alignment has also been tackled using linguistic knowledge, such as bilingual dictionaries, thesauri, POS taggers etc. (Ker and Chang 1997, Pianta and Bentivogli 2004). Hybrid methods combine statistical and linguistic knowledge. For instance, the SIMR/GSA system (Melamed 2000) establishes points of correspondence based on several criteria and groups them into 'chains'. Criteria include cognateness (similar words in source and target language, like English *calibrate* and French *calibrer*) and bilingual lexical entries. Gelbukh and Sidorov (2006) compare source and target paragraphs through a bilingual dictionary.

2.4. Word alignment for translators

The principle of word alignment discussed above is of great interest for application in a translation environment. However, from a specialised translator's point of view, we have to take into account that efficiency is of utmost importance. Therefore, a word alignment approach can only be useful for translators if it complies with a number of criteria. Firstly, it should be applicable to parallel texts (e.g. translation memories) of various sizes, especially to texts of restricted size. Secondly, the word alignment should be fine-grained and accurate enough in order for the translator to detect a word's or word group's translation in context. However, there is no need for the alignment to treat each single word in the text, as specialised translators are primarily interested in content words like nouns. Thirdly, the approach should minimise the number of user parameters and intermediate steps in order to increase time effectiveness.

3. Hypothesis

Language independence is an important strength of the statistical approach to word alignment. However, this approach shows some specific features to be taken into account. Firstly, the performance of statistical word alignment systems like GIZA++ (Och and Ney 2003) increases with the size of the parallel text. Secondly, parameter values need to be optimised depending on the alignment task. Thirdly, statistical systems consider words as surface strings, which may cause problems if word forms are associated with the same lemma or multiwords are linked to

single words, etc. Due to these features, statistical methods have been combined with linguistic knowledge (see section 2.3). Though tackling a number of issues present in purely statistical methods, they require fine-tuning of parameter values and preprocessing steps.

This article starts from the assumption that a certain amount of linguistic knowledge is a prerequisite in a flexible alignment framework. Such a framework is independent of the text size, requires minimal user input and preprocessing, and links elements below the surface form. As such, the framework complies with the needs of a professional translator environment as described in 2.4. More specifically, we start from the assumption that basic forms of linguistic data, such as dictionaries and glossaries, are a sufficient basis for a word alignment method that provides the degree of flexibility, accuracy and information required in that environment. In the next section, we describe our approach in more detail.

4. Method: A linguistics-based approach

4.1. Linguistic resources

We propose a method that makes use of linguistic resources which can be easily manipulated or extracted from richer data: bilingual dictionaries consisting of lemmas and their translations, bilingual glossaries, lists linking word forms to their lemmas, and stop word lists. The use of stop word lists is motivated by the fact that we are primarily interested in content words, and the alignment of function words is more complicated.
The above resources provide strong evidence for linking source and target words or word groups. As the presence of cognates also provides such evidence, our method includes a mechanism for the comparison of words.

4.2. Alignment procedure

4.2.1. Text parts

Our alignment procedure splits the source and target texts in a number of parts of an approximate length (e.g. about 2,000 words). The source and target text are split into an equal number of parts and each text part starts and ends with a full sentence.

Each source text part is aligned both with the corresponding target text part and its immediate context. For instance, the tenth part of the source text is compared with the sequence of the ninth, tenth and eleventh target text part. We refer to this sequence as the *target text sequence*. We prefer this approach to the procedure of performing sentence alignment before starting word alignment as the latter procedure may be hampered by large divergences between the source and the target text.

The alignment of a source-text part with its target text sequence takes place in two steps. The first step (see section 4.2.2) creates candidate links between small text units, called *fragments*, from the source text part and its target text sequence. In the second step (see section 4.2.3), a number of candidate fragment links is grouped into *link paths*, which align larger text areas such as a sentence part, sentence or paragraph.

When all source-text parts have been aligned with their target text sequence, an alignment is produced for the whole source and target texts by gathering all the link paths created for all the source text parts and grouping the link paths into larger units called *metapaths* (see section 4.2.4).

4.2.2. Fragment links

The first step in aligning a source text part with its target text sequence is the creation of candidate fragment links. Fragments are words, word groups and word parts occurring at specific positions in the text, hence having a specific context. For instance, if the same word occurs twice in a text, we are dealing with two different fragments. We proceed through the following steps:

1. We create a set of source fragments, consisting of the following elements present in the source text part: words (except for stop words), word groups of a limited length and word parts. We restrict the latter to word prefixes with a minimal length. For instance, we take the following prefixes in consideration for the word *emballage*: *emballag*, *emballa*, *emball* and *embal*. In the same manner, we create a set of target fragments from the target text sequence.

2. We create a number of *variants* for each source and target fragment: lemmas of words and word groups (using the word form lists), and alternative spellings of a word (e.g. removal of hyphens in word groups).

3. We link a source and a target fragment if they or their variants are equal (cognateness, see example 2) or if a dictionary or glossary entry exists with a source lemma identical to the source fragment or a variant, and with a target lemma identical to the target fragment or a variant (see example 3). We also create a link if a prefix of a lemma is equal to a fragment (see example 4). So, a part of a given word in the text may be linked to a part of a word in a dictionary or glossary.

(2) FRA *HP4598*
NLD *HP-4598*

(3) FRA *magasins* de musique
NLD muziek*winkels*
ENG 'music shops'
LEX *magasin → winkel*

(4) FRA pendant l'*emball*age des produits
NLD terwijl we produkten *verpakken*
ENG 'while we pack products'
LEX *emballer → verpakken*

4. Links whose source and target fragment is completely covered (subsumed) by the source or target fragment of another link are discarded. For instance, the link in example 5 makes the link in example 3 redundant.

(5) FRA *magasins de musique*
NLD *muziekwinkels*
ENG 'music shops'
LEX *magasin de musique → muziekwinkel*

Fragment links have the potential to bridge small divergences between source and target text, such as categorial differences, differences in grammatical features, and translation equivalence between a word and a part of a composed word (see section 2.2). However, it is important to note that the first alignment step produces *candidate* fragment links. These may include links between words in unrelated sentences.

4.2.3. Link paths

The second step in aligning a source text part and its target text sequence consists of selecting and grouping a number of candidate fragment links in order to build *link paths* spanning a text area of the source and target text. This step aims at aligning a piece of the source text with a piece of the target text that represents its translation. Link paths are created in two steps:

1. We determine sequences of fragment links complying with the following criteria:

   - The distance between the source fragments, in terms of the number of characters, should not exceed a predefined maximum (vicinity rule). Idem for target fragments.
   - There is no overlap between the source fragments. The same holds for target fragments.

- A sequence should not be overlapped by a sequence that covers a larger area of the source and target text. If so, the sequence is discarded (non-overlap rule).

Example 6 presents a sequence of fragment links.

(6) FRA *conduire*$_1$ une voiture de *sport*$_2$ est *dangereux*$_3$
NLD *rijden*$_1$ met een *sport*$_2$wagen is *gevaarlijk*$_3$
ENG 'driving a sports car is dangerous'

2. Within the sequences of fragment links determined in the previous step, we create locally crossing links by adding new fragment links; see example 6, in which a new link *voiture* → *wagen* follows the second link in the source text and precedes it in the target text. The distance between the fragments of a new link and the fragments of the link that is being crossed should not exceed a maximum number of characters.

Obviously, it is always possible that a link path contains two irrelevant fragment links that coincidentally occur near one another. However, the more links a link path contains, while complying with the criterial rules of vicinity, non-overlap and local crossing, the lower the odds that the link path as such is completely irrelevant, i.e. that two text areas are linked that are not each other's translation.

The source and target areas covered by a link path can have an arbitrary size, ranging from a sentence part to a number of subsequent sentences. Link paths bridge certain types of divergences described in section 2.2, such as the equivalence between a word group and a composition and diverging word order.

4.2.4. Metapaths

We gather all link paths created for all source text parts and we group the link paths into metapaths using the same procedure as for grouping fragment links into link paths. We adopt the criterial rules of vicinity, non-overlap and local crossing. The maximal distances in terms of character number, however, are much larger than for grouping fragment links. The non-overlap rule concerns link paths created for adjacent source text parts; the latter's target text sequences overlap. If two link paths overlap, the one covering the least source and target text area is discarded.

A metapath bridges divergences such as sentence parts occurring in a different order in the source and target texts, or the omission of a paragraph during translation (see section 2.2). The final word alignment of the source and target text consists of a number of metapaths; more specifically, the alignment consists of a number of fragment links grouped into a number of link paths, which are in turn grouped into a number of metapaths.

4.3. Implementation

We implemented our approach using TAWK, a commercial version of AWK (Thompson Automation Software 1996). Our system is called FragmALex, which stands for *Fragment Alignment using a Lexicon*. It runs as a command-line program on the Microsoft Windows platform, but it should be possible to adapt it for UNIX and Linux by using the appropriate TAWK compiler. The program outputs all fragment links present in the final word alignment. The output consists of lines with four fields: (1) the position of the first character of the source text fragment, (2) the length of the source text fragment, in number of characters, (3) the position of the first character of the target text fragment, and (4) the length of the target text fragment. The program also produces a few files containing a graphical view of the alignment (see image 1 in section 6.1).

4.4. Evaluation procedure

4.4.1. Formal evaluation

In order to be able to evaluate the system output in a formal way, we devised a procedure based on recall and precision for comparing the set of fragment links created by the system with another set of fragment links. The latter set is either generated by another configuration of our system (if we want to compare different configurations) or a manually created gold standard; see Ahrenberg, Merkel, Hein and Tiedemann (2000). In order to cope with the fact that not only full words or word groups are compared, but also parts of words, we determine for each fragment link in the system output whether it partially or fully overlaps in both the source and target texts with a link in the other set. We refer to this type of equivalence between links as *loose* equivalence, as opposed to *strict* equivalence, which requires a link to be identical to another one.

We calculate the precision of the system output to another link set by dividing the number of system output links equivalent to a link in the other set by the total number of system output links (0 <= precision <= 1). In order to calculate the recall (completeness of a link set with respect to another link set), we divide by the total number of links in the other set (0 <= recall <= 1). Recall and precision either involve loose equivalences or strict equivalences.

The precision and recall measure can be adopted when the FragmALex output is compared with a gold standard. The recall formula can also be employed for comparing system configurations that use specific linguistic resources. If we compare a configuration A with a configuration B that makes use of an additional linguistic resource, we expect the recall to be higher than if we compare system B with A; in other words, we expect the linguistic resource to have an added value. We wrote a script in order to automate comparisons.

4.4.2. Gold standard

In order to be able to evaluate the system output by means of a gold standard, we devised a method for manually aligning fragment links. The first step is the alignment of the sentences of the parallel text. If a source sentence corresponds to multiple target sentences or vice versa, we perform a one-to-many alignment. The second step consists in the alignment of fragments in the source and target sentences, by labelling the fragments with a numerical identifier (see example 7 in section 5.3). A fragment labelled with the number 0 has no translation equivalent.

We limited the scope of the fragment alignment for two reasons. First of all, the task of creating a gold standard for word alignment is notoriously difficult and subjectivity cannot be avoided (Melamed 1998). Secondly, our system is designed for applicability in a translation environment rather than for an exhaustive alignment. For these reasons, we restricted the alignment to words or word groups that belong to open classes (e.g., auxiliary verbs are excluded), we restricted the alignment to continuous word groups, we mapped source fragments to a single target fragment and vice versa, and we ignored content mismatches, such as paraphrases. Syntactic mismatches, such as the translation of a verb by a noun, were included in the alignment.

We created a script converting the manual annotation into the four-field format of the FragmALex output. Obviously, a comparison of the manual annotation with the system output is only valid if the system input consists of the aligned sentences rather than running source and target text.

5. Case study: medical texts

5.1. Corpus

We compiled a parallel corpus consisting of medical articles in French and Dutch. We chose the language pair French-Dutch for two reasons. Firstly, the multilingual Belgian context provides a high volume of translated material in these two languages. Secondly, both languages are rather divergent in terms of linguistic features and wording, which presents an interesting challenge for the automation of word alignment. For the medical community, a great number of magazines covering various disciplines are available online. For our purposes, we compiled a corpus of articles from online magazines on neurology and cardiosurgery of the Belgian medical publisher *Transmed* (*Transmed* 2006), which appear both in Dutch and French. The original articles are either written in French or Dutch. The magazines address domain specialists and, as a consequence, show specific features of LSP (Language for Special Purposes), like the frequent occurrence of domain-specific terminology and nominalizations.

The publisher permitted us to access to the *Transmed* website, which is normally only available to certified medical practitioners through a login.

In our selection of bilingual articles for inclusion in the corpus, we applied a number of criteria. The first criterion concerned text length. We ensured that the sample contained a variety of article lengths (see requirement in section 2.4). The second criterion was textual coherence. Articles consisting of lists of abstracts were not included in the sample. The third criterion consisted of the degree of text complexity. Articles in the sample had to be domain-specific, though not too specialised, because this would increase the presence of cognates and numbers, thus making the alignment too obvious.

We kept the number of articles in the corpus limited, in order to make the study feasible. The corpus consists of 18 articles. Each of them being available as a web page on the *Transmed* website, we downloaded them and removed extraneous text such as web page headers and bibliographical references at the end of the text. The lengths of the articles range from 230 to 2,800 words, with an average of 1,300 words.

5.2. Linguistic resources

We compiled the following linguistic resources: a bilingual medical glossary, a general-purpose bilingual dictionary, word form lists for French and Dutch, and stop word lists for both languages.

We created the bilingual medical glossary on the basis of an online glossary of technical and popular medical terms developed by medical specialists and linguists (Vander Stichele 1995), which contains 1830 concepts in up to nine European languages. As an example, the concept 0088 (labelled *analeptic* after the English word*)* is expressed in Dutch by the term *analepticum* (technical variant) or *versterkend middel* (popular variant). We downloaded all web pages containing the concepts in order to retrieve the French and Dutch terms belonging to the same concept and to convert them into bilingual entries. Each entry contains a unique combination of a French and Dutch term (lemma), such as *analeptique* → *analepticum* and *stimulant* → *analepticum*. The collection of about 2,000 entries resulting from the above procedure constitutes our medical glossary.

In order to create the general-purpose bilingual dictionary, we linked all French and Dutch lemmas from the general-purpose, machine-readable dictionary Ergane, developed by Gerard van Wilgen (TL Online Inc. 2007), which contains Microsoft Access tables with lemmas from a considerable number of languages, among which Esperanto. The lemmas in the other languages are linked to the numerical identifier(s) of the equivalent Esperanto lemma(s). We downloaded the table for Dutch (the developer's mother tongue), containing about 61,000 unique lemmas, and the French

table of about 11,000 unique lemmas. We wrote a script that generates entries containing a combination of a French and Dutch lemma that have the same Esperanto identifier. This resulted in a bilingual dictionary with about 29,000 entries. For instance, the French word *tissu* and the Dutch words *stof* en *weefsel* (English 'tissue') share an Esperanto identifier, which leads to two entries, *tissu* → *stof* and *tissu* → *weefsel*. As *stof* also has other Esperanto identifiers, for meanings such as 'subject' and 'stuff', it appears in additional entries.

As for the creation of the French and Dutch word form lists that link word forms to their lemma, we used two resources. For Dutch, we used the WebCelex site (Max Planck Institute for Psycholinguistics 2001), which provides online access to the CELEX lexical databases for Dutch, English and German created by the Centre for Lexical Information of the University of Nijmegen. We created the Dutch word form list by generating a subset of lexical data in WebCelex (the resulting entries include the fields *Orthography→Word* and *Lemmas→Lemma Head*) and running a script that filters out duplicate entries (identical word forms with different grammatical features) and entries in which the word form and the lemma are identical (redundant information). For French, we created a script that retrieves and filters word forms and lemmas from the French flat text file that is part of the MULTEXT set (Véronis 1998). The final Dutch word form list contains about 207,000 entries; the French list about 192,000.

We created our final resource manually. It consists of stop word lists, with function words such as pronominal words, frequent prepositions and auxiliary verbs. The stop word lists contain a few hundred words. They allow us to focus our alignment procedure on content words.

5.3. Gold standard

For two of the *Transmed* articles from the corpus (one on brain research[2], one on spinal cord injury[3]), we created a gold standard. The size of these French and Dutch articles ranges between 1,700 and 2,000 words. Example 7 shows a sample sentence from the gold standard.

(7) FRA Cela *signifie*$_1$-t-il *aussi*$_2$ que des *troubles*$_3$ du *sommeil*$_4$ peuvent *influencer*$_5$ cette *consolidation*$_6$ du *processus*$_7$ d'*apprentissage*$_8$?
NLD *Betekent*$_1$ dit *ook*$_2$ dat *slaap*$_4$*stoornissen*$_3$ deze *consolidatie*$_6$ van het *leer*$_8$*proces*$_7$ *negatief*$_0$ kunnen *beïnvloeden*$_5$?
ENG 'Does this also mean that sleep problems may influence the consolidation of the learning process ?'

6. Results

6.1. Manual assessment

We ran FragmALex using all resources described in section 5.2 on the 18 articles of the medical corpus, using French as the source language, and manually assessed arbitrary samples of the output. When we observe undershoot or incorrectness in the system's output, it appears to be caused by the incompleteness of the lexical data, paraphrases, and short target text spans containing multiple fragments that are lexically equivalent to the same source word. The system partially tackles the incompleteness problem through the comparison of a word or word part in the source or target text with the prefix of a lemma in the glossary or dictionary. This is illustrated by the fragment link in example 8. Comparing word parts is also useful for non-identical cognates and for words with different syntactic categories.

(8) FRA *respectifs*
NLD *respectie*velijke
ENG 'respective'
LEX *respectif* → *respectief*

As for the creation of metapaths, image 1 illustrates a divergence in the order of the subject (French *le débit sanguin cérébral régional* and Dutch *het regionale cerebrale bloeddebiet*) and other sentence constituents (French *pendant la nuit … H2150* and Dutch *tijdens de studienacht ... techniek*). FragmALex captures this divergence by creating two link paths (paths 15 and 16 in the image) and grouping them into a metapath, in which they cross each other locally. The lower part of the image presents the fragment links contained in link path 15.

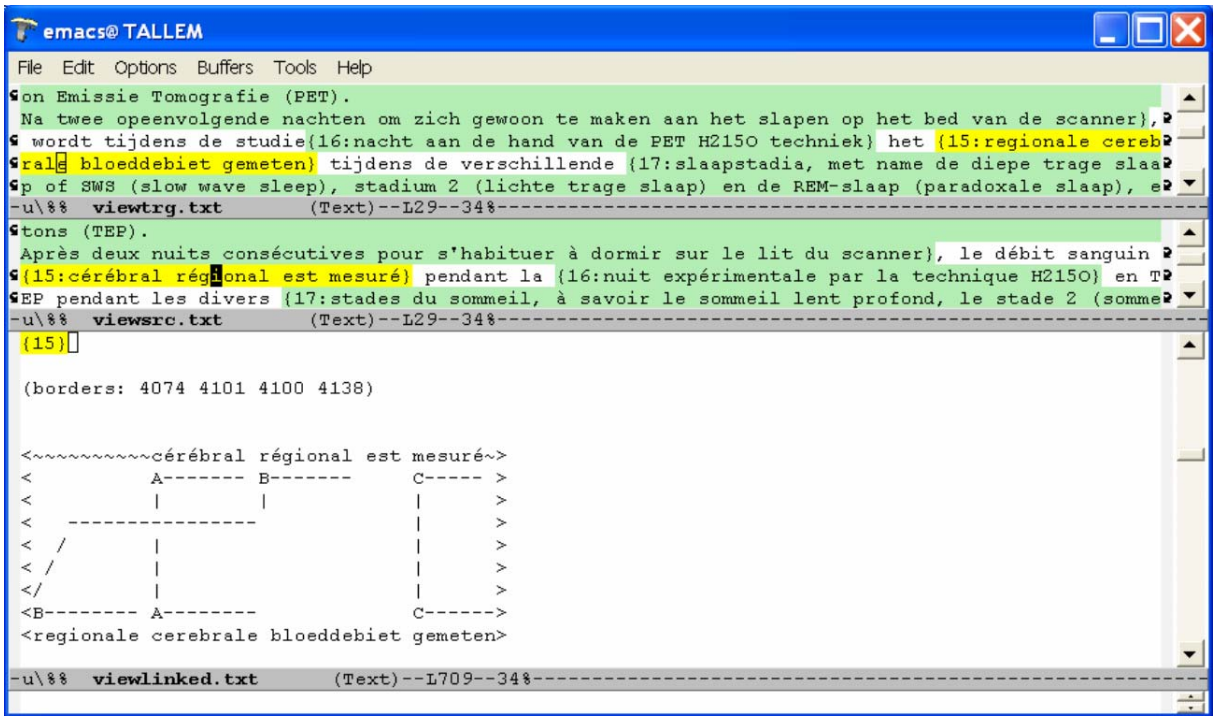

Image 1: FragmALex output: crossing link paths within a metapath.

6.2. Evaluation of resource addition

In order to determine the added value of linguistic resources, we ran FragmALex in three different configurations on the 18 articles in the medical corpus. The first configuration only recognises cognates and makes no use of linguistic resources. The second configuration recognises cognates and employs the bilingual medical glossary, and the third configuration uses, in addition, the general-purpose bilingual dictionary. Both the word form and stop word lists were used consistently in each of the three configurations. Table 1 presents the recall of the output of a configuration (left column) with respect to the output of a configuration that differs by a single linguistic resource (top row). The first number in each cell indicates the recall involving *loose* equivalences, the number between parentheses the *strict* recall.

Configuration	**Cognateness**	**+ Glossary**	**+ Dictionary**
Cognateness		0.87 (0.66)	
+ Glossary	0.97 (0.74)		0.59 (0.52)
+ Dictionary		0.95 (0.83)	

Table 1: FragmALex recall for configurations differing by a single resource.

It is apparent from the loose recall figures in table 1 that the first configuration produces a lot of the fragment links also found in the output of the second configuration: 87% of the links of the second configuration appear in the first one. Vice versa, almost all links produced by the second configuration appear in the output of the first one (97% of all links in the first configuration). Adding a glossary to FragmALex has some added

value since it increases the number of fragment links. The increase is, however, much more spectacular if the general-purpose bilingual dictionary is added to the system: only 59% of the fragment links in the third configuration appear in the second configuration. The strict recall figures in the table are lower in absolute terms, but are interrelated in the same way as the loose ones.

The general-purpose bilingual dictionary has a higher added value than the bilingual medical glossary because the articles under consideration use a specialised language, which prefers scientific terms to their popular alternatives. E.g., Dutch *fractuur* ('fracture') is preferred to its popular alternative *beenbreuk* ('broken bone'). As scientific terms tend to be similar across languages, they can also be linked as cognates (the Dutch word *fractuur* resembles French *fracture*).

6.3. Gold standard evaluation

We submitted the aligned sentences from the two articles in the gold standard to FragmALex. We ran the system in three configurations (see previous section). Table 2 lists the loose recall and precision of the system's fragment links with respect to the fragment links in the gold standard. The strict figures appear between parentheses.

Article 1	**Recall**	**Precision**
Cognateness	0.26 (0.05)	1.00 (0.20)
+ Glossary	0.28 (0.10)	0.99 (0.34)
+ Dictionary	0.65 (0.43)	0.93 (0.62)
Article 2		
Cognateness	0.37 (0.09)	0.99 (0.25)
+ Glossary	0.43 (0.19)	0.98 (0.43)
+ Dictionary	0.64 (0.39)	0.92 (0.56)

Table 2: FragmaLex recall and precision for gold standard.

The loose recall increases with the addition of lexical information, in line with our expectations. Recall figures become consistent across both articles if both the glossary and the dictionary are used. We observe a tendency towards a slight precision decrease when adding linguistic resources. This can be explained by the fact that an increase in lexical information causes FragmALex to generate more potential fragment links, hence more alternatives. However, the mechanism creating link paths and metapaths ensures that the increase in recall is much higher than the decrease in precision when linguistic resources are added.

As for the strict figures, we observe the same increase in recall. However, the precision increases rather than decreases. This can be explained by the fact that cognate recognition involves a lot of similar rather than identical words, and thus mostly concerns non-identical links.

7. Conclusions and future research

In this paper, we introduced a word alignment approach geared towards specialised translators. Commonly available systems assisting translators support the creation and retrieval of terminological information in a rather time-consuming way. Methods for word alignment, being non-trivial because of linguistic and extralinguistic divergences between the source and target text, require preprocessing steps and fine-tuning of parameter values. Statistical methods are language independent but depend on text size.

Our approach, implemented as a program called FragmALex, primarily focuses on efficiency and usability. It focuses on content words, which are of primary interest to specialised translators, and omits the requirement for preparatory sentence alignment. It combines low-structured linguistic knowledge with a mechanism for linking fragments, including word parts, and grouping them into link paths and metapaths, thus filtering out irrelevant fragment links. This approach allows us to capture divergences, such as differences in syntactic categories, and, on a larger text level, differences in wording and text layout. The program output presents the resulting fragment links within their context. This allows the translator to rapidly detect a translation equivalent or spot the equivalent of an unlinked word through its position near a fragment link. Parameter tuning is restricted to a minimum; it concerns, for instance, the maximal distance between fragment links or between link paths.

We applied FragmALex to a medical corpus in French and Dutch that we compiled from articles of restricted size. We compiled linguistic resources such as a medical glossary and a general-purpose bilingual dictionary. A comparison of different system configurations reveals that the dictionary has a higher added value than the glossary because the specialised language of the corpus focuses on terms that may be present in the glossary but can as well be linked as cognates. Comparing the system output with a gold standard using *loose* equivalences shows that the mechanism creating link paths and metapaths prevents the addition of lexical resources from decreasing the precision to a great extent. As a manual assessment of the system shows, the mechanism linking word parts is able to prevent the system from producing certain incorrect fragment links when lexical evidence is lacking.

In future research, we will compare the performance of our approach with programs such as GIZA++ and SIMR. We also project the usage of larger bilingual dictionaries, which is not straightforward due to copyright issues, and of multilingual thesauri, which allow for the identification of target words that are not full-fledged translation equivalents but belong to the same semantic field; see Ker and Chang (1997). We will perform extraction of specialised terms not present in the lexical resources by exploiting part-of-speech information and associating source and target

words that appear near fragment links. Such a procedure employs existing lexical data in order to create new ones. Finally, in order to increase the usability of the system, we will convert the program output into a web page displaying the fragment links as hyperlinks within running text.

References

- Ahrenberg, Lars, Magnus Merkel, Anna Såvall Hein and Jörg Tiedemann (2000). "Evaluation of word alignment systems." *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 1255–1261.

- Bowker, Lynne (2002). *Computer-Aided Translation Technology: A Practical Introduction.* Ottawa: University of Ottawa Press.

- Brown, Peter, Vincent Della Pietra, Stephen Della Pietra and Robert Mercer, (1993). "The mathematics of statistical machine translation: parameter estimation." *Computational Linguistics* 19(2), 263–311.

- Gale, William and Kenneth Church (1991). "A program for aligning sentences in bilingual corpora." *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 177–184.

- Gelbukh, Alexander and Grigori Sidorov (2006). "Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming." *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition*, Cancún, Mexico, 824–833.

- Kay, Martin and Martin Röscheisen (1993). "Text-translation alignment." *Computational Linguistics* 19(3), 121–142.

- Ker, Sue and Jason Chang, J. S. (1997). "A class-based approach to word alignment." *Computational Linguistics* 23(2), 313–343.

- Max Planck Institute for Psycholinguistics (2001), *WebCelex.* Online at: http://www.mpi.nl/world/celex (consulted on 10.09.2006).

- Melamed, I. Dan (1998). Manual annotation of translational equivalence: the Blinker project, *Technical Report #98-07*, Institute for Research in Cognitive Science.

- Melamed, I. Dan (2000). "Pattern recognition for mapping bitext correspondence." J. Véronis (ed.). *Parallel text processing. Alignment and use of translation corpora.*, Kluwer Academic Publishers, 25–47.

- Och, Franz Josef and Ney, Hermann (2000b). "Improved statistical alignment models." *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, 440–447.

- Och, Franz Josef and Ney, H. (2003). "A systematic comparison of various statistical alignment models." *Computational Linguistics* 29(1), 19–51.

- Pianta, Emanuele and Bentivogli, Luisa (2004). "Knowledge intensive word alignment with KNOWA." *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland, 1086–1092.

- Planas, Emnanuel (2000). "Extending translation memories." *Proceedings of the 5th EAMT Workshop*, Ljubljana, Slovenia, 7-19.

- Ricca, Filippo, Paolo Tonella, Emanuele Pianta and Christian Girardi (2004). "Experimental results on the alignment of multilingual web sites." *Proceedings of the 8th European Conference on Software Maintenance and Reengineering*, Tampere, Finland, 288–295.

- Thompson Automation Software (1996). *TAWK version 5.0.* Online at: http://www.tasoft.com (consulted on 03.01.2006).

- TL Online Inc. (2007). *Ergane version 8.0.* Online at: http://www.travlang.com/Ergane (consulted on 07.01.2007).

- *Transmed* (2006). Online at: http://www.transmed.be (consulted on 08.09.2006).

- Vander Stichele, Robert (1995). *Multilingual glossary of technical and popular medical terms in nine European languages. Technical report*. Heymans Institute of Pharmacology, University of Gent and Mercator College, Department of Applied Linguistics. Online at: http://users.ugent.be/~rvdstich/eugloss/welcome.html (consulted on 12.10.2006).

- Véronis, Jean (1998). *Multext-Lexicons. A set of Electronic Lexicons for European Languages.* CD-ROM, distributed by ELRA/ELDA.

Biographies

Tom Vanallemeersch has been working in language and speech technology for over ten years. Currently, he is coordinating the translators' skills lab at the Applied Linguistics department of Lessius University College in Antwerp, Belgium. His PhD focuses on automated alignment and term extraction.

Dr. Cornelia Wermuth specialises in terminology and teaches medical translation at Lessius University College. She wrote a PhD on frame-based representation of medical classification themes.

Contact: tom.vanallemeersch@lessius.eu

[1] In the examples, we use the language codes FRA, NLD and ENG for French, Dutch and English sentences. The linked French and Dutch text elements are marked in italics. If necessary, we distinguish links by numbering them. The English sentence is a literal translation of the French and Dutch sentences. If the latter have a different meaning, two English translations are specified. In some examples, the lexical information used for linking text elements is included in a LEX field for the sake of clarity.

[2] The French article titled *Le renforcement de la mémoire spatiale dans l'hippocampe humain pendant le sommeil lent profond* and its Dutch equivalent, from the *Neuropsy* magazine, Number 36 (2004).

[3] The French article titled *Problématique du patient vieillissant atteint de lésion médullaire: complications musculosquelettiques et neurologiques* and its Dutch equivalent, from the *Spectrum 50* magazine, Number 3 (2005)