Translation alignment and lexical correspondences : a methodological reflection

Olivier Kraif

▶ **To cite this version:**

Olivier Kraif. Translation alignment and lexical correspondences : a methodological reflection. Granger, Sylviane and Altenberg, B. Lexis in Contrast, Benjamins Publisher, pp.271–290, 2001. hal-01073722

HAL Id: hal-01073722
https://hal.science/hal-01073722

Submitted on 30 Sep 2018

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 alignment and lexical correspondences: a methodological reflection

Olivier Kraif

1. Introduction

In the last few years much interest has been given to the outcome of translation aligning: Isabelle (1992) proposed using bilingual parallel texts, or bi-texts, i.e. segmented and aligned translation corpora, as a *Corporate Memory* for translators. He alleged that "existing translations contain more solutions to more translation problems than any other existing resource". Such a translation database, organised as a bilingual concordancer (as in the TransSearch Project, cf. Simard *et al.* 1993) would store all the previously found solutions for a given translation problem and allow the translator to recover them easily. Other alignment-based tools, such as automatic verification, have a natural place in a translator's workstation. Error detection can be implemented when translations are provided in aligned format. In the TransCheck system, Macklovitch (1995a) shows how common errors such as "deceptive cognates, calques, illicit borrowings" can be automatically detected in a bi-text framework. Other features, such as exhaustiveness (i.e. omission errors; cf. Isabelle *et al.* 1993) or terminological consistency (Macklovitch 1995 b), can be tested. It is also possible to verify automatically, in a reliable manner, the proper translation of specific phrasal constructions such as dates or numerical expressions. The transduction grammar formalism seems to work very well in this kind of restricted translation task.

In the more ambitious field of Example-Based Machine Translation (Sato & Nagao 1990, Brown *et al.* 1990), aligned corpora form the cornerstone of the system. The linguistic knowledge is stored implicitly in the recorded examples of translation. The success of the system depends on the huge quantity of aligned sentences that constitute mutual translations.

Another interesting application is the automatic extraction of bilingual lexicons. Many works (Dunning 1993, Dagan *et al*. 1993, Gaussier & Langé 1995) have shown how to use statistical filters to pair lexical units that have a similar distribution in each part of the bi-text. As a large proportion of these similar units are translation equivalents, they can be useful in establishing bilingual (or multilingual) glossaries for empirical observation.

In order to align parallel texts, several techniques have been implemented which have yielded satisfactory results. Even when they take advantage of lexical information most of the systems work at sentence level (Brown *et al.* 1991, Simard *et al*. 1992, Kay & Röscheisen 1993, Gale & Church 1991). Indeed, it is a well-known fact that the hypothesis of parallelism does not hold below sentence level, and 'lexical alignment' appears to be a far more complex problem. However, some systems have yielded encouraging results in producing lexical alignment (Brown *et al.* 1993).

Given the huge variety of algorithms and techniques devoted to alignment, we are now entering an evaluation phase, and some large-scale projects such as Arcade (Langlais *et al*. 1998) set out to give a coherent framework for definition and evaluation of the aligning task. In the former project two different tasks have been tested: *sentence alignment* and *lexical spotting* (i.e. finding lexical correspondences for a given list of test words). The evaluation task consists of two steps: given a test corpus, we have to determine first a *gold standard*, i.e. a manually constructed alignment that is considered to be exact. Then we have to implement a *metric* in order to effect a quantitative comparison of any other alignment with the standard. Both in the case of sentence and of word track, two kinds of difficulty resulted from the definition of a standard alignment: segmentation discrepancy and correspondence problems.

Detailed criteria were given to human aligners and annotators in order to cope with inconsistencies, but the lexical spotting task, in respect of sentence alignment, rapidly proves problematic.

After giving a precise definition of what bilingual alignment involves, we will go on to describe various problems associated with alignment at word level. We will then show the inconsistency of such a concept, and draw a line between the extraction of lexical correspondences and the alignment task from a general point of view. We believe that only a proper definition of the concepts of alignment and correspondence that takes account of the actual practice of translation can produce reliable criteria for the creation of a gold standard that can be used for the purpose of evaluation.

2. The concept of alignment

The standard concept of alignment can be summed up as follows:

> *Aligning consists in finding correspondences, in bilingual parallel corpora, between textual segments that are translation equivalents.*

Translation equivalence is above all a global property of the translation of a text. It is not a linguistic property, but a pragmatic one: the translation arrived at is a result of interpretative choices that are made in a specific situational context. As Sager (1994: 186) says:

> While the cognitive and linguistic equivalents are mainly established at the level of the sentence or in smaller units during the translation phase, the pragmatic equivalents have to be selected first in the preparation phase and at the level of the text type before being also realised in smaller units at appropriate points in the document.

These extra-linguistic parameters are linked to many factors at the pragmatic level: text typology, text intention, receptors, dynamic equivalence (cf. Nida & Taber 1982), cultural adaptation, conceptual background and so on.[1]

Translation equivalence is a relationship between messages entrenched in two given contexts and backgrounds: the source and the target context. This global equivalence does not imply equivalence at the level of linguistic units. In the following example, the original advertisement for golf items is not translated at word level (Henry 1991: 15):

(1) To make your greens come true
Pour faire putt de velours

The French version includes a pun, as in English: it refers to the expression *faire patte de velours*, which means ‘to sheathe its claws’ (of a cat). *Putt* is a particular stroke in golf, and the translation plays on the paronymy between ‘putt’ and *patte*.

This example illustrates the fact that the equivalence holds at a global and an abstract level. The two versions ‘work’ in the same way, although using different linguistic means. In this case the relevant features are the pun and the theme. Depending on the function of the message, some features are more relevant than others, and have to be maintained in translation whatever the cost (while other features are lost): these may be the conceptual content or rhetorical figures, stylistic devices, formal features such as alliteration, and so on.

Therefore, to segment and to establish correspondence between segments, we have to make a specific assumption about the translation. We might call it *translational compositionality*. This concept is developed by Isabelle (1992):

> For translation to be possible at all, translational equivalence must be *compositional* in some sense ; that is, the translation of a text must be a function of the translation of its parts, down to the level of some finite number of primitive equivalences (say between words and phrase).

I do not completely agree with Isabelle when he presents compositionality as a condition of the possibility of translation. Compositionality may be a characteristic of the *process* of translation, but remains a relative notion as as far as the *product* of translation is concerned. In fact, the translational compositionality of a bilingual corpus determines exactly the level at which it is possible to align it.

In more formal terms, the compositionality assumption leads to the definition of a specific corpus structure: the bi-text. Generally speaking, a bi-text is a quadruple <T1,T2,Fs,A> where T1 and T2 are mutual translations (the direction of the translation is irrelevant), Fs is a segmentation function which divides the texts into a set of smaller units (e.g. paragraphs, sentences, phrases), and A is the alignment of these units, i.e. a subset of the product Fs(T1) x Fs(T2).

This general definition can lead to different kinds of bi-text: Fs can produce either a complete or a fragmentary partition of the texts, or a hierarchical partition where different levels are simultaneously involved (paragraph, sentence, words). Moreover, we can focus on particular alignments with several restrictions. For instance, Isabelle & Simard (1996) define a *monotone* alignment in terms of three constraints:

- *no crossing correspondences*: i.e. the segments must appear in the same order in both texts.
- *no partially overlapping segments*: two different segments that appear in different pairings cannot share the same portion of text. For instance, the phrase *Machine Aided Translation* would not yield two segments: *Machine Aided* and *Aided Translation*.
- *no discontinuous correspondences*: i.e. there are no discontinuous segments, such as *Machine* […] *Translation* in the previous example.

Most existing alignment systems use this kind of monotone alignment. Indeed, in the current state of the art, the possibility of automatic alignment is strongly conditioned by the *parallelism* of the corpora. As Gaussier & Langé (1995: 71) have defined it, parallelism consists in the conjunction of two criteria: *one-to-one matching* and *monotony*:

- *One-to-one matching* means that each segment of one text has a correspondence in the other text. In fact, this condition is never completely realised, because translation induces additions and omissions. Therefore, this criterion is *more or less* satisfied, depending on the particularities of the translation.
- *Monotony*, as previously defined, is also a relative property. In general, however, inversion of the sequence of segments is rare.

3. Alignment techniques

As Simard & Plamondon (1996) point out, alignment techniques can produce two different kinds of result:

- *alignment* involving a parallel segmentation of both texts into smaller logical units (such as paragraphs, sentences or even phrases), in such a way that the n^{th} segment of source text and the n^{th} segment of target text are mutual translations.
- a *bi-text map* involving a set of points (x,y), called *anchor points*, where x and y refer to precise locations in the source and the target text that denote portions of text corresponding to one another.

The latter case is very general, because it does not presuppose a previous segmentation. But a bi-text map is not a very useful form of bi-text, as it does not directly indicate correspondences between textual units as in bilingual concordances: it only establishes connections between text areas. We consider the bi-text map as a preliminary and intermediate step for the achievement of a full alignment.

In the following discussion, I will give examples of sentence alignment, but the problems are the same for every kind of segmentation compatible with compositionality.

What is alignment?

Bilingual alignment is not a negligible problem, as translation does not preserve unit boundaries. Practically, a sentence can be translated by two or more sentences, or can simply be omitted. At every stage the alignment algorithm has to determine the appropriate clustering of units in order to respect the translation equivalence property. We can illustrate this by the example in the following table, extracted from an English translation of Jules Verne's novel *De la terre à la lune* (which is a part of the BAF corpus, developed at the CITI of Montreal, which has been used as a benchmark in the Arcade Project; cf. Langlais *et al*. 1998 and Simard 1998: 489).

Table 1: *Example of sentence alignment*

English text		French text	
P1	"Here we are at the 10th of August," exclaimed J.T. Maston one morning, "only four months to the 1st of December.	P'1	! *" Nous voilà au 10 août, dit un matin J.-T. Maston*
		P'2	*Quatre mois à peine nous séparent du premier décembre !*
		P'3	*Enlever le moule intérieur, calibrer l'âme de la pièce, charger la Columbiad, tout cela est à faire !*
P2	We shall never be ready in time !"	P'4	*Nous ne serons pas prêts ! "*

We can write this alignment as follows:

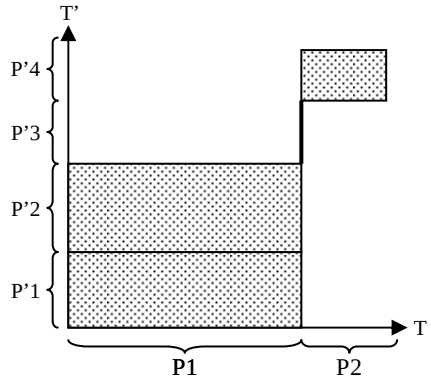
T=P1P2 T'=P'1P'2P'3P'4 A= {[P1;P'1P'2],[∅;P'3],[P2;P'4]}

It is also possible to represent these clusters as a sequence of *n-p* transitions, called an *alignment path*:

A = (1-2), (0-1), (1-1)

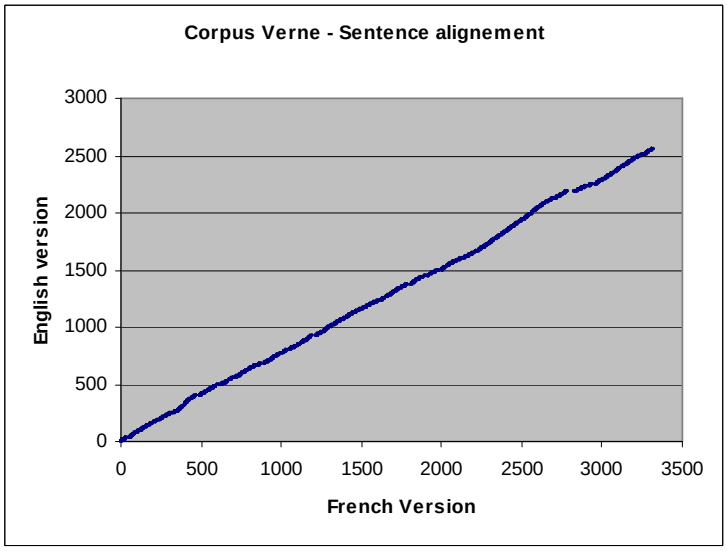
Figure 1 gives a two-dimensional representation of this path, with T and T' on the X and Y axes. The alignment is represented by the surfaces involved in the segment pairings:

Figure 1 *: bidimensional representation of an alignment*



If we draw a chart representing the complete translation of Verne's novel, we get a general view of the path, as shown in Figure 2.

Figure 2 : *a complete alignment path*



The more parallel the translation is, the closer the path is to the diagonal of the square.

General framework

Several methods have been developed to calculate this kind of path automatically. They are usually implemented within a probabilistic framework: by estimating the probability of all possible paths, the algorithm can find the best-scoring one, i.e. the one with the highest probability.

Given a function p(A) which estimates the probability of alignment A, the algorithm has to find:

$A^* = \text{argmax}_A\ p(A)$

Naturally, this task of maximisation creates great problems of computation: the number of possible paths is in O(n!) (where n represents the number of sentences). A Viterbi algorithm which considers simultaneously all the sub-paths that share the same beginning can reduce the computation to $O(n^2)$ but it is still a considerable problem.

A simpler method of reducing search space is to consider only the paths that are not too far from the diagonal. This is a direct implication of the parallelism hypothesis: if omissions, additions and inversions are marginal, the path cannot diverge too much from the diagonal.

Prealignment

Another way of reducing search space is a preliminary extraction of a rough but reliable bi-text map, based on superficial clues. Chapter separators, titles, headers and sometimes paragraph markers can yield information of great interest to produce a quick and acceptable pre-alignment (Gale & Church 1991). Other superficial clues are the chains that remain invariant in translation, such as proper nouns or numbers (Gaussier & Langé 1995). If one had to align a text and its translation manually in a completely unknown language, one would use exactly the same superficial, straightforward information. I have shown elsewhere (Kraif 1999) that such chains can be used to align 20% to 50% of the different texts in the BAF corpus (with less than 1% error rate).

Alignment clues

Once the search space has been reduced, we can evaluate the probability of each possible sentence cluster in order to calculate the global probabilities of each path. Different kinds of information are available for this estimation.

Segment length

Gale & Church (1991) and Brown *et al*. (1991) simultaneously developed a length-based method which yielded good results on the Canadian Hansard Corpus.[2] The principle of this method is very simple: a long segment will probably be translated by a long segment in the target language, and a short segment by a short one. Indeed, Gale & Church show empirically that the ratio of the source and target lengths corresponds approximately to a *normal distribution*. Note that it is possible to compute the segment lengths in two ways: as the number of characters or the number of words in the segment. According to Gale & Church, the length in characters seems to be a little more reliable in the case of translations between English and French (the variance of the ratio is slightly smaller). Using the average and the variance of this ratio as specific parameters, depending on the language pairs involved, they compute the probability of a cluster as a combination of two factors: the probability of length ratio and the probability of transition. These latter probabilities were determined in an empirical way in the case of the Gale & Church corpus, considering only six of the most frequent types of transition, viz.:

One sentence – one sentence :	p(1-1)=0.89
One sentence – zero sentence and reciprocally :	p(1-0)=p(0-1)=0.0099
Two sentences – one sentence and reciprocally :	p(2-1)=p(1-2)=0.089
Two sentences – two sentences :	p(2-2)=0.011

All the other alignment clues are based on the lexical content of the segment. They come from a very straightforward heuristic: word pairings can lead to segment pairings. If two segments are translation equivalents, they will probably include more lexical units that are translation equivalents than any independent segments would. To take the lexical information into account, one just needs to know which units are potential equivalents. This linguistic knowledge can be extracted from various sources including bilingual dictionaries and bilingual corpora.

Bilingual dictionaries

To be usable for this purpose, dictionaries have to be available in electronic format. Moreover, in technical fields, it is not always easy to find a dictionary that is consistent with the corpus concerned.

Bilingual corpora

It is also possible to extract a list of lexical equivalents directly from a bilingual corpus. Indeed, translation equivalents usually have very similar distributions in both texts. These distributions can be converted into a mathematical form and then be compared quantitatively. In the K-vec method, developed by Fung & Church (1994), both texts are divided into K equal segments. Then, for each word (here the words are treated as lexical units), it is possible to compute a vector representing its occurrence in each segment: with 1 for the i^{th} co-ordinate if the word appears in the i^{th} segment, otherwise 0. Thus, when both words have 1 for the same co-ordinate, one can say that they co-occur. This model of co-occurrence (cf. Melamed 1998) makes it possible to calculate the similarity of two distributions by several measures based on probabilities and information theory.

In two texts divided in N segments, for two words W1 and W2 occurring in each text in N_1 and N_2 segments respectively, and co-occurring in N_{12} segments, you can easily compute their mutual information:

$$I = \log \frac{\frac{N_{12}}{N}}{\frac{N_1}{N} \cdot \frac{N_2}{N}}$$

If N_1 and N_2 are not too small (>3), then beyond a certain threshold of mutual information (I>2), it is highly improbable that the N_{12} co-occurrences are due to chance: you can assume that they are linked by a special contrastive relation, which may be translational equivalence. For rarer events (N_1 or $N_2 \leq 3$), other measures, such as the likelihood ratio (Dunning 1993) or the t-score (Fung & Church 1994), are more suitable.

The problem of the K-vec method is that segments are big (because the system has no knowledge about the real sentence alignment) and the co-occurrences model is very imprecise. The finer the alignment, the more exact the word pairing obtained.

As there is an interrelation between segment pairing and word pairing, some systems work in an iterative framework (Kay & Röscheinsen 1993, Débili & Sammouda 1992). From a rough prealignment of the corpus they extract a list of word correspondences. From these correspondences they then compute a finer alignment. From this new alignment they extract a new and more complete set of word pairings. And so on, until the alignment has reached stability.

Formal resemblance

Another way of determining lexical equivalence is to focus on cognate words which share common etymological roots, such as the French word *correspondance* and the English word *correspondence*. Cognateness is defined by Simard *et al*. (1992) as word pairs which share the

same first four characters (*4-grams*), including also invariant chains such as proper nouns and numbers. Simard *et al.* show empirically that cognateness is strongly correlated with translation equivalence. On the basis of a probabilistic model, they estimate the probability of a segment cluster given its cognateness. This model, combined with the length-based model, yielded significant improvements of the results achieved by Gale & Church. In previous works, we show that a special filtering of cognate words can give a very precise and complete prealignment: in the case of the BAF corpus, we obtained 80% of the full alignment, with a very low error rate (about 0.5%). Of course, the exploitation of formal similarities depends on the languages involved. In the case of related languages such as English and French, cognateness is important. In the case of technical texts we can expect to observe cognates even between unrelated languages, because technical and scientific terms usually share common Graeco-Latin roots.

4. The concept of lexical correspondence

Usually, lexical correspondences are treated as a particular case of alignment. In the Arcade project, for instance, lexical spotting is seen as a simpler sub-problem of full alignment. Brown *et al.* (1990) give the following example of what can be described as word alignment:

(2) The poor don't have any money
Les pauvres sont démunis
A={(The ; *Les*) (poor ; *pauvres*) (don't have any money ; *sont démunis*)}

Even if it is generally admitted that the condition of quasi-monotony does not hold in this case, the supposed one-to-one matching seems to justify the concept of word alignment. Let us examine the problems that are involved here.

Segmentation discrepancy

From a monolingual point of view, a lexical unit is defined in terms of syntactic and semantic autonomy. A compound expression can be characterised by the conjunction of several criteria:

- a certain degree of semantic non-compositionality.
- more or less syntactically frozen structure.
- a certain recurrence.

We will not discuss the complexity of this problem. The definition of a lexical unit is a difficult problem in linguistics, and no consensus has been reached so far in the linguistic community.

In any case, it appears that the units emerging from lexical alignment do not have lexical consistency, depending only on the structural homology between the related segments. For instance, another translation of the previous sentence results in different units:

(3) The poor don't have any money
Les pauvres n'ont pas d'argent
A={(The ; *Les*) (poor ; *pauvres*) (don't have ; *n'ont pas*) (any ; *d'*) (money ; *argent*)}

Lexical alignment yields non-lexical compounds, but it can also break up genuine lexical units. For example, we can align the English, French and Italian expressions in different ways:

(4) To be the very devil
Avoir le diable au corps
Avere il diavolo in corpo
French/Italian: A ={(*Avoir* ; *Avere*) (*le* ; *il*) (*diable* ; *diavolo*) (*au* ; *in*) (*corps* ; *corpo*)}
English / French: A = {(To be the very devil ; *Avoir le diable au corps*)}

In this case we have word-for-word correspondence inside the lexical unit across Italian and French. The problem is: should the lexical alignment be allowed to break up lexical compounds, when it is possible?

Semantic discrepancies

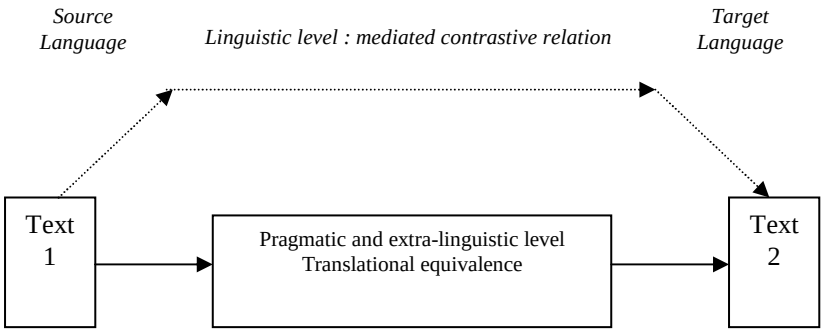
Another problem is semantic discrepancy, which is common between a text and its translation. The following example is extracted from a European Parliament report.[3]

(5) the marking of banknotes for the benefit of the blind and partially sighted
l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite
[literally: 'the issue of banknotes identifiable by the blind and partially sighted persons']

The phenomenon of semantic discrepancy is frequently found in the practice of translating. This can be explained by the importance of the extra-linguistic level. Translation, as Pergnier notes (1993: 23), is not only an operation between two different languages, it is first a transformation between *messages*, involving the whole pragmatic and conceptual background.[4] As Pergnier (1993: 75) says, "the equivalence at both levels, between two utterances and between the signs that they include, does not exist before the translation, but is a consequence of it" [my translation].

Thus the contrastive level, i.e. the possible equivalence between signs of different systems, is secondary: it is a result of translation as an act of communication, as shown in Figure 3.

Figure 3 : *the level of translational equivalence*



As a result, lexical alignment based on semantic criteria is very often unclear. In these two sentences

(6) the various policies for access to employment for disabled people

les différentes politiques mises en œuvre pour permettre l'accès des personnes handicapées à l'emploi
[literally: 'the various policies implemented to allow disabled people to access a job']

divergent solutions are possible for the following phrases:

A={(for ; mises en œuvre pour permettre)}

or else, if we take omissions into account:

A={(∅ *; mises en œuvre*) (for ; *pour*) (∅ *; permettre*)}

These semantic discrepancies, combined with segmentation difficulties, create very complex configurations in lexical alignment. Consider the following case:

(7) The assessment of the official cause of death is a piece of information vital to these registers.
Pour la bonne tenue de ces registres, l'évaluation des cas de mortalité constatés par les autorités apporte des informations importantes.
[literally: 'For the good keeping of these registers, the evaluation of causes of death noted by the authorities gives important information']

In these sentences we observe correspondences between discontinuous units:

A={(vital ; *importantes […] pour la bonne tenue de ces registres*)}

There are thus two possible alignments of the following phrases:

A={(cause of death ; *cas de mortalité*) (official ; *constatés par les autorités*)}

or

A={(official cause of death ; *cas de mortalité constatés par les autorités*)}

Since semantic discrepancy and segmentation inconsistency are not discrete phenomena, but follow a continuum of intensity, the determination of reliable criteria to solve this kind of alignment is almost impossible.

Recently great attention has been given to automatically extracted bilingual glossaries. Indeed, as we have seen before, probabilistic models make it possible to extract lexical correspondences by comparing the distribution of lexical items in a parallel corpus. Large-scale evaluations, as in the Arcade project, have been designed to test these methods and to guide the construction of a gold standard, established on the basis of a test corpus, in order to benchmark the different systems. In order to cope with the problems inherent in the concept of lexical alignment and delineate more clearly the task of automatic lexical pairing, we propose a redefinition of the concept of *lexical correspondence*.

Lexical correspondences

We agree with Debili (1997: 200) that lexical alignment is "neither one-to-one, nor sequential, nor compact. Correspondences are fuzzy and contextual." He therefore proposes to distinguish between "lexical correspondence", where the mutual translation can be validated by a bilingual dictionary, and "contextual correspondence" (1997:203), i.e. translation that depends on a specific context. But we do not subscribe to this point of view. The attestation of a dictionary is a somewhat arbitrary criterion, and it does not reflect the inherent continuity of the phenomena.

We prefer to distinguish two different kinds of task: alignment and the determination of correspondences. Indeed, *lexical correspondence* can be defined in a very restricted sense:

> *A **lexical correspondence** is a relation of denotational (conceptual, extra-linguistic) equivalence between two lexical units in the context of two segments that are translation equivalents.*

This definition raises the following issues:

- lexical units are linguistically defined, in a *monolingual* context. By adopting a broad definition of lexical units, including compounds, phraseology and even terms, it is possible to avoid the issue of segmentation inconsistency. If the problem is shifted to a monolingual point of view, its resolution appears to be far more reasonable.
- we focus on the contextual sense of the lexical unit (referring to the opposition between "*signe type*" and "*signe occurrence*" made by Rastier 1991: 96).
- monotony and one-to-one matching are no longer presumed, in accordance with empirical observations.

We feel that lexical alignment is a nebulous notion which inherits most of the misleading statements from the first generation of MT systems. For instance, in this case:

(8) the marking of banknotes for the benefit of the blind and partially sighted
l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite.

We can draw the following correspondences:

> C={(banknotes ; *billets de banque*) (blind ; *aveugles*) (partially sighted ; *personnes à vision réduite*)}

The rest of the sentences is just a normal translation residue, due to the divergences between the two versions. These divergences can have a linguistic cause (e.g. morphosyntactic or lexical differences) or not (e.g. conceptual inferences).

Maximal resolution alignment

This kind of lexical correspondence differs from sub-sentence alignment. We define a special kind of alignment that is very often confused with lexical correspondence:

> *A **maximal resolution alignment** is a matching of the smallest possible segments in accordance with the principle of translational compositionality.*

This kind of alignment does respect the criteria of parallelism, except for monotony below the sentence level. In such an alignment, the syntactic characterisation of the segments is not determined: it can be a word, a phrase, a whole sentence, or even a paragraph. This depends on whether the translation is literal or not: if the translation of a sentence cannot be decomposed, the sentence has to be considered as a whole.

Translation spotting, as defined in the Arcade project, appears to be a kind of maximal alignment, and yet it is fragmentary: it focuses on segments that contain some specific lexical units. For instance, looking for the correspondence of the French word *apporter*, it yields the alignment between the boldfaced segments:

(9) A meeting held in Brussels […] went a long way towards **meeting the concerns** expressed by the Honourable Member.
Une réunion, qui s'est tenue à Bruxelles […] a permis d'accentuer l'effort pour ***apporter des éléments concrets de réponse aux préoccupations*** *exprimées par l'honorable parlementaire.*

The notions of *translational compositionality* and *maximality* capture very neatly the criteria of translation spotting. In discussions about the appropriateness of aligning *peas* with *pois* in the phrases *green peas* and *petits pois*, the non-compositionality of this translation pair gives a very clear solution: *petits pois* and *green peas* cannot be decomposed.

The characteristics of lexical correspondence and maximal alignment are summed up in Table 2.

Table 2: *Characteristics of Lexical Correspondence and Maximal Alignment*

	Lexical Correspondence	**Maximal Alignment**
Segmentation criterion	Monolingual, lexical unit level	Segmentation depends on structural homology between texts. It is based on both translational compositionality and on maximality: the segments cannot be decomposed further.
Formal characteristics	Usually one-to-one relations between some lexical units, and the rest is residual. Many-to-many relations are also possible.	Quasi-bijection, quasi-monotony below sentence level.
Syntactic nature of the segments	Lexical unit: words, compounds, set phrases, terms.	No syntactic consistency: word, phrase, sentence, paragraph.
Pairing criterion	Denotational identity (in the occurrence context).	Translation equivalence

To illustrate these two concepts, we give another example:

(10) Confidential secret service information on applicants for European civil service posts
Récolte de données à caractère personnel par les services secrets d ' un État membre sur les candidats aux concours organisés par les institutions européennes

The maximal alignment could be as follows:

> A={(Confidential *; à caractère personnel*) (secret service: *par les services secrets*) (∅ ; *d'un État membre*) (information ; *Récolte de données*) (on ; *sur*) (applicants ; *les candidats*) (for European civil service post ; *aux concours organisés par les institutions européennes*)}

And we can extract the following lexical correspondences:

> C={(confidential ; *personnel*) (secret service ; *services secrets*) (information ; *données*) (on ; *sur*) (applicant ; *candidat*) (European ; *européennes*)}

5. Conclusion

These reflections aim at defining and clarifying the key concepts of alignment and correspondence in the field of bi-text exploitation and evaluation. We make a distinction between two different types of bilingual pairing: the alignment of the smallest segments that are considered as translational equivalents (in accordance with the principle of translational compositionality), and the lexical correspondences which concern stable lexical units (in a broad sense) having the same denotational content. In fact, inside two aligned sentences, there is no need to have all lexical units correspond with each other. Semantic discrepancies between a sentence and its translation can be very important, and the assumption of quasi-bijection does not hold at the lexical level.

This distinction opens up a number of new possibilities:

- the development of more consistent criteria in order to establish benchmark corpora in the field of evaluation,
- a more accurate interpretation of the meaning of contrastive phenomena which emerge from a bi-text. The sets of textual segments constituting a bi-text are not linked by specific linguistic properties, but by translational equivalence, which is defined at an extra-linguistic level. Of course, contrastive regularities can be observed at different levels: morpho-syntax, lexicology, terminology and phraseology. But these regularities are not rules: they emerge statistically from the recurrence of translation facts.

Notes

1. “Dynamic equivalence is therefore to be defined in terms of the degree to which receptors of the message in the receptor language respond to it in substantially the same manner as the receptors in the source language.” (Nida and Taber 1982:24).
2. The Canadian Hansard Copus consists in a French / English Canadian Parliamentary Proceedings, available at http://www.parl.gc.ca/36/1/parlbus/chambus/house/debates/indexe/homepage.html
3. These reports can be found at http://www.europarl.eu.int
4. “Dire que la traduction opère sur des messages, c’est en effet proclamer qu’elle est un acte de communication (ou d’échange linguistique) avant d’être un acte de comparaison inter-linguale.” (Pergnier, 1993:23)

Acknowledgements

Many thanks to Kim Van den Broecke, Hélène Ledouble and Luc Bardolph for their helpful assistance in the editing of this article.

References

Brown, P., Cocke, J., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. and Roossin, P. 1990. "A statistical approach to machine translation". *Computational Linguistics* 16: 79-85.

Brown, P., Della Pietra, S. and Mercer, R. 1993. "The mathematics of statistical machine translation: parameter estimation". *Computational Linguistics* 19: 263-311.

Brown, P., Lai, J. and Mercer, R. 1991. "Aligning sentences in parallel corpora". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 169-176. Berkeley, CA.

Dagan, I., Church, K.W. and Gale, W. 1993. "Robust bilingual word alignment for machine aided translation". In *Proceedings of the Workshop on Very Large Corpora, Academic and Industrial Perspectives*, 1-8.

Debili, F. 1997. "L'appariement: quels problèmes?". In *Actes des 1ère JST FRANCIL de l'AUPELF UREF*, 199-206. Avignon.

Debili, F. and Sammouda, E. 1992. "Appariements de phrases de textes bilingues Français-Anglais et Français-Arabes". In *Actes de COLING-92*, 528-524. Nantes.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics* 19: 61-74.

Fung, P. and Church, K.W. 1994. "K-vec: A new approach for aligning parallel texts". In *Proceedings of the 15th International Conference on Computational Linguistics*, 1096-1102. Kyoto.

Gale, W. and Church, K.W. 1991. "A program for aligning sentences in bilingual corpora". In *Proceedings of the 29th Annual Meeting of the ACL*, 177-184. Berkeley, CA.

Gaussier, E. and Langé, J.-M. 1995. "Modèles statistiques pour l'extraction de lexiques bilingues". *T.A.L.* 36 (1-2): 133-155.

Isabelle, P. 1992. "La bi-textualité: vers une nouvelle génération d'aides à la traduction et la terminologie". *Meta* XXXVII (4): 721-731.

Isabelle, P., Dymetman, M., Foster, G., Jutras, J.M. and Macklovitch, E. 1993. "Translation analysis and translation automation". In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in MT*. Kyoto.

Israël, F. and Lederer, M. 1991. *La liberté en traduction. Actes du colloque international tenu à l'E.S.I.T. les 7,8 et 9 juin 90.* Paris. Didier Erudition, Coll. traductologie.

Kay, M., Röscheisen, M. 1993. "Text-translation alignment". *Computational Linguistics* 19: 121-142.

Kraif, O. 1999. "Identification des cognats et alignement bi-textuel: une étude empirique". In *Actes de la 6ème conférence annuelle sur la Traitement Automatique des Langues Naturelles. TALN 99*, 205-214. Cargèse, France.

Langé, J.-M. and Gaussier, E. 1995. "Alignement de corpus multilingues au niveau des phrases". *T.A.L.* 36 (1-2): 133-155.

Langlais, Ph., Simard, M. and Veronis, J. 1998. "Methods and practical issues in evaluating alignment techniques". In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics.* Montréal, Canada.

Macklovitch, E. 1995a. "Can terminological consistency be validated automatically?". In *Proceedings of the IVèmes Journées scientifiques, lexicommatiques et dictionnairiques, organized by Aupelf-Uref.* Lyon, France.

Macklovitch, E. 1995b. "The future of MT is now, and Bar-Hillel was (almost entirely) right". Centre d'innovation en technologies de l'information (CITI). Laval, Canada. [Available at http://www-rali.iro.umontreal.ca.]

Melamed, I.D. 1998. "Models of co-occurrence". In *Technical Report #98-05.* Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA. [Available at http://www.cis.upenn.edu/~melamed/home.html]

Nida, E.A. and Taber, C.R. 1982. *The Theory and Practice of Translation*. Leiden: Brill.

Pergnier, M. 1993. *Les fondements socio-linguistiques de la traduction.* Lille: Presses Universitaires de Lille.

Rastier, F. 1989. *Sens et textualité*. Paris: Hachette, Coll. HU.

Sager, J.C. 1994. *Language Engineering and Translation: Consequences of Automation*. Amsterdam: John Benjamins.

Sato, S. and Nagao, M. 1990. “Towards memory-based translation”. In *Proceedings of COLING’90*, 247-252. Helsinki.

Simard, M., Foster, G. and Isabelle, P. 1992. “Using cognates to align sentences”. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 67-81. Montréal, Canada.

Simard, M., Foster, F. and Perrault, F. 1993. “TransSearch: un concordancier bilingue”. Centre d’innovation en technologies de l’information (CITI), Laval, Canada. [Available at URL http://www-rali.iro.umontreal.ca.]

Simard, M. 1998. “The BAF: a corpus of English-French bitext”. In *Proceedings of First International Conference on Language Resources and Evaluation*, 489-494. Granada, Spain.