From Translational Data to Contrastive Knowledge: Using Bi-text for Bilingual Lexicons Extraction

Olivier Kraif

Title : From Translational Data to Contrastive Knowledge: Using Bi-text for Bilingual Lexicons Extraction

Author: Olivier Kraif

Affiliation: LIDILEM – Université Grenoble 3 - Stendhal

Address : BP 25 - 38040 Grenoble Cedex 9 – France

http://www.u-grenoble3.fr/kraif

E-mail :Olivier.Kraif@u-grenoble3.fr

Abstract

Textual aligning consists in pairing segments (e.g. sentences or phrases) that are translational equivalent across corpora of translations. An interesting application of textual aligning is the automatic extraction of bilingual lexicons. As it has been pointed out during previous evaluation campaigns, such as Arcade, lexical aligning remains problematic. In order to solve problems of consistency linked with the concept of translational compositionality, a redefinition of lexical aligning task is proposed, introducing the concept of lexical correspondence. Simple techniques dedicated to lexical correspondences extraction are then evaluated. Thus, it appears that adapted statistical filters allow to extract very accurately significant regularities that are relevant at the contrastive level. More generally, these methods prove to be adapted not only for bilingual lexicons extraction: they could be used to study a wide range of contrastive phenomena on empirical basis.

Keywords: bi-text, alignment, lexical correspondence, translation, contrastive linguistics.

1. Introduction

In the last few years, much interest has been given to the outcome of translation aligning : Isabelle (1992) proposed to use bilingual parallel texts, or *bi-texts*, i.e. segmented and aligned translation corpora, as a “corporate memory” for translators. In that kind of corpora, the linguistic and translational knowledge is implicitly stored in the recorded examples of translation.

The Arcade Project (Véronis & Langlais 2000), through a large scale evaluation of aligning systems, demonstrated that sentence aligning was already a mastered technology, for most parallel corpora. However, with the *translation spotting* [1] task evaluated during the second campaign, in 1998, lexical aligning proved to be far more difficult.

An interesting application of that kind of aligning is the automatic extraction of bilingual lexicons. A lot of works (Dunning 1993, Dagan *et al*. 1993, Gaussier & Langé 1995, Boutsis & Piperidis 1996, Melamed 1998a, Kraif 2000) 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 great proportion of these similar units are translational equivalents, they can be useful to establish bilingual (or multilingual) glossaries upon empirical observation.

The first section addresses the problems inherent to the lexical alignment concept, and shows they are partly due to a lack of consistency in its definition. To cope with this difficulty, and delineate more clearly the task of automatic lexical pairing, a redefinition of the concept of *lexical correspondence* is proposed. The implementation of simple techniques, based on lexical distributions and cognateness, is then described. The results are evaluated according to a manually extracted set of lexical correspondences. It appears that an objective criterion, the conditional entropy, is strongly correlated to the quality of the output.

Finally, the relative part of speech distributions for the corresponding units are compared. This basic example illustrates how these techniques can be generalised to make richer observations concerning any contrastive phenomena, and compare different languages upon rigorous empirical basis.

2. Aligning at lexical level, a problematic task

To align a parallel corpus, i.e. to segment and to pair corresponding segments, we have to make a specific assumption about the translation. It can be called *translational compositionality*. This concept was developed by Isabelle (1992: 3):

> 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 part, down to the level of some finite number of primitive equivalences (say between words and phrases).

Thus, compositionality is a relative property, which closely depends on the scope of these primitive equivalences. The local degree of compositionality will finally determine the granularity of the bilingual alignment.

In more formal terms, the compositionality assumption leads to the definition of a specific corpora structure: the bi-text. Isabelle (1992: 4) gives the following definition: a bi-text is a quadruple <T1,T2,Fs,A> where T1 and T2 are mutual translations (we do not take the direction of translation into account), 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 fragmentary partition of the texts, or a hierarchical partition where different levels are simultaneously involved (paragraphs, sentences, words). Moreover, we can focus on par-

ticular alignments with several restrictions. Most of the existing aligning systems deal with *monotone* alignment (Isabelle & Simard 1996), where the segments must appear in the same order in both texts. Indeed, in the current state of the art, the possibility of aligning automatically is strongly conditioned by the *parallelism* of the corpora. As Gaussier & Langé (1995: 71) have defined it, parallelism consists of 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. This criterion is the formal expression of compositionality. In fact, this condition is never completely realised, because translation induces additions and omissions. Therefore, this criterion is *more or less* met, depending of the specificities of the translation.

- *Monotony*, i.e. the stability of the order of the translated segments, is also a relative property. Usually, inversions in the sequence of segments are marginal.

For a large variety of corpora, these two criteria generally hold for small clusters of sentences. For instance, in the BAF corpus (Simard 1998) involved in the Arcade Project, including technical, scientific, legal and institutional French-English texts, only a very small part of the corpus (an alphabetically sorted glossary) was not monotonous at sentence level.

Of course, monotony does not stand at word level, as noticed by Gaussier & Langé (1995), because of the syntactic differences between languages. Yet, lexical alignment is commonly presented as a particular case of alignment. Brown *et al* (1993: 267) give the following example of what can be defined as “word alignment”:

(1) Eng.: The poor don’t have any money
Fr.: Les pauvres sont démunis
A={(The ; Les) (poor ; pauvres) (don’t have any money ; sont démunis)}

The supposed one-to-one matching seems to justify the concept of word alignment. But as we showed in another discussion (Kraif 2001), this criterion raises two related problems:

2.1 Segmentation inconsistency

The term “word alignment” is misleading because, most of the time, it is not possible to align *single* words, but *clusters* of words.. In the example above, “don’t have any money” is neither a word, nor a compound word. Units that are yielded by this kind of pairing have no linguistic consistency: they just depend on specific choices of the translator. For instance, another translation of the previous sentence results in different units:

(2) Fr.: The poor don’t have any money
Eng.: Les pauvres n’ont pas d’argent
A={(The ; Les) (poor ; pauvres) (don’t have ; n’ont pas) (any ; d’) (money ; argent)}

2.2 Semantic discrepancy

Another problem is due to semantic variations that commonly occur between a text and its translation. The following example is extracted from the JOC corpus used in the Arcade Project.

(3) Eng.: Illegal transactions involving the heritage
Fr.: Transactions illégales aux dépens du patrimoine

Should we align *involving* with *aux dépens du* (literally *at the expense of*)? The semantic connection between the two expressions is rather fuzzy, because the French expression is more specific than the English one. Outside the particular context of the translation above, the pair (*involving* ; *aux dépens du*) has no clear meaning. Debili (1997 : 203) points out that there are

two kinds of correspondence : “lexical correspondences”, that could be found in a bilingual dictionary, and “contextual correspondences” that rely on a “local and contextual construction, based on a ‘’human understanding’’ of the two sentences”(« recomposition locale et contextuelle, fondée sur une “ compréhension humaine ” des deux phrases »).

Indeed, the example above shows that the translational equivalence does not automatically imply the semantic equivalence of the words that are involved. Translational equivalence is the result of choices made by the translator, which depend on the purpose of the communication, and which are linked with a lot of factors at the pragmatic level : textual typology, text intention, receptors, cultural adaptation, conceptual background, etc. When Nida (1969 : 14) gave his own definition of translating, he was very close to St Jerome’s opposition *ad verbum* / *ad sensum* (cf. Letter LVII to Pammachius on the best method of translating): “translating consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style.” But here, *meaning* has to be interpreted according to the communicational background, and it can be set at different levels that are more or less salient according to the intention of the message. For instance, at the pragmatic level, Nida (1969 : 142) introduces what he calls *dynamic equivalence*: “Dynamic equivalence is therefore to be defined in terms of the degree to which receptors of the message in the target language respond to it in substantially the same manner as the receptors in the source language.” Of course, there are other levels of equivalence linked with other functions of the message: conceptual or referential, metalinguistic, poetic, rhetorical, etc.

The translational equivalence is a relation between *messages* rooted in two different contexts and backgrounds. Thus, local linguistic structures must give way to the global changes required by the adaptation, as the means are subordinate to the goal. In the following example,

given by Jacqueline Henry in Israël & Lederer (1991: 15), the original advertisement for golf items is not translated at word level:

(4) Eng.: To make your greens come true
Fr.: 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 one's claws' for a cat. *Putt* is a particular stroke in golf, and the translation plays on the similarity between *putt* and *patte*. 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 (while other features are lost).

Even the need for conceptual equivalence can lead to linguistic transformations. Martin Kay (2000: xiii) gave the following example, found in the scientific literature :

(5) Eng. : Gravity is a pervasive force in the world… (*Scientific American*)
Fr. : La pesanteur s'exerce partout sur la terre… (*Pour la science*)
[literally : Gravity applies everywhere on earth]

There is a semantic link between (Eng.) *pervasive* and (Fr.) *partout*, that would allow to potentially align them, but it raises the problem of the utility of such an alignment outside this particular context. As Kay (2000: xiv) said: "For a researcher interested in high-quality translation, an alignment program that paired *pervasive force*, or at least *pervasive*, with *partout* (everywhere) might stimulate important insights, but as a source of potential entries in a bilingual dictionary, it might constitute a source of frustration."

These semantic discrepancies follow a continuum of intensity. Combined with the segmentation inconsistency, they are the source of two major problems concerning lexical aligning:

- It is difficult to draw a line between omissions (or additions) and normal semantic differences. In these two sentences divergent solutions are possible:

(6) Eng.: the various policies for access to employment for disabled people
Fr.: 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 to a job]

If we accept semantic discrepancies, we have:

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

Or else, if we accept omissions, the alignment is :

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

- Variations in segmentation are strongly linked with variations in semantic equivalence. In many occurrences, a finer granularity of the alignment results in a less satisfactory equivalence at the linguistic level. Observe the following case:

(7) Eng.: The assessment of the official cause of death is a piece of information vital to these registers.
Fr.: 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]

There are different possible alignments for the following phrases:

A1={(official cause of death ; cas de mortalité constatés par les autorités)}
A2={(cause of death ; cas de mortalité) (official ; constatés par les autorités)}

A3={(cause ; cas) (of ; de) (death ; mortalité) (∅ ; constatés) (∅ ; par) (∅ ; les) (official ; autorités)}

In the first alignment, the extracted phrases can be used as translational equivalents in other contexts: taken as a whole, they have a close semantic interpretation. On the other side, the third alignment is more fine-grained, but some pairs are not semantically equivalent (e.g. *official* and *autorités*). The second alignment is an intermediate configuration. How to make a choice between these different solutions?

Since semantic discrepancy and segmentation inconsistency are not discrete phenomena, it is very difficult to find reliable criteria in order to cope with arbitrary choices.

2.3 Parallel commutation

M.-D. Mahimon (1999 : 34) proposed an original test to determinate a lexical alignment in a more consistent way. She suggested to implement the Catford's (1965 : 28) concept of *commutation* :

> "In place of *asking* for equivalents we may adopt a more formal procedure, namely, *commutation* and observation of concomitant variation. In other words we may systematically introduce changes into the SL [source language] text and observe what changes if any occur in the TL [target language] text as a consequence. A *textual translation equivalent* is thus : *that portion of a TL text which is changed when and only when a given portion of the SL text is changed.*"

A similar idea underlies the method developed by Malavazos *et al*. (2000:1.2) for the extraction of translation templates:

> "The main idea is based on the observation that given any source and target language sentence pair, any alteration of the source sentence will most likely result in one or more changes in the respective target sentence, while it is also highly likely that constant and variable units of the source sentence correspond to constant and variable target units respectively."

In the classic commutation test, linguistic units such as phonemes are induced from the parallel commutation of form and meaning: a phonetic variation become significant if it implies a semantic variation. The bilingual commutation test described by Mahimon concerns a source and a target sentence, and involves both directions:

1. From form to meaning: by commuting a unit of the source text, a difference of meaning is produced with the target text.

2. From meaning to form: in the target text, this semantic difference must be cancelled by commuting some target units, in order to restore the translational equivalence.

According to Mahimon, such a test allows to align source and target lexical units, by pairing every group of units that switch in the same time. She gave the following example (1999: 41):

(8) Fr.: **Ce** projet de loi prévoira un système de déclaration des maladies infectieuses
Eng.: **This** bill will provide for an infectious disease notification system

When we switch *Ce* with *Chaque* the equivalence can be restored by replacing *This* with *Each* :

(9) Fr.: **Chaque** projet de loi prévoira un système de déclaration des maladies infectieuses
Eng.: **Each** bill will provide for an infectious disease notification system

To refer to this parallel commutation, we write : *Ce* || *This*

Thus, we can align *Ce* with *This*.

In her definition, Mahimon gave different criteria to guarantee a minimal granularity and a full translational compositionality. We can systematise these criteria with the two following principles:

- *Minimal commutation* : « when possible, commutation must concern one word (or morpheme) at once. » (1999 : 36) The switching parts should be as small as possible.

- *Transitivity*. Units that switch together in the same side (source or target) constitute equivalence class. We introduce two other relations $\equiv_s$ et $\equiv_t$ with the following definition:

$$\exists\ U' \,/\, U_1 \parallel U' \text{ et } U_2 \parallel U' \ \Leftrightarrow U_1 \equiv_s U_2$$

$$\exists\ U \,/\, U \parallel U_1' \text{ et } U \parallel U_2' \ \Leftrightarrow U'_1 \equiv_t U'_2$$

where U_1, U_2, U are units of the source sentence and U_1', U_2', U' units of the target sentence. The equivalence class closure is obtained by transitivity:

$U_1 \equiv_s U_2$ and $U_2 \equiv_s U_3 \Rightarrow U_1 \equiv_s U_3$

$U_1' \equiv_t U_2'$ and $U_2' \equiv_t U_3' \Rightarrow U_1' \equiv_t U_3'$

Both relations $\equiv_s$ and $\equiv_t$ are equivalence relation in a mathematical sense : they are reflexive, commutative and transitive. It has to be noted that they are monolingual relations : they result in clusters of units in a same language. They should not be confused with the transitive relations established by Simard (2000: 53).

Then, the commutation relation can easily be extended to these clusters. We have:

$C \parallel C' \Leftrightarrow \exists\ U \in C, U' \in C' \,/\, U \parallel U'$

where C is a cluster of units of the source sentence, and C' a cluster of units of the target sentence.

Mahimon (1999 : 43) gave the following examples, formatted according to our own conventions (original units are in bold face, switching units are in normal style) :

(10) Eng.: This bill will (**provide for** / confirm) an infectious disease notification system

Fr .: Ce projet de loi (**prévoira** / entérinera) un système de déclaration des maladies infectieuses

$\Rightarrow$ *provide* $\equiv_s$ *for* || *prévoira*

(11) Eng.: This bill (**will provide** / provides) for an infectious disease notification system

Fr.: Ce projet de loi (**prévoira** / prévoit) un système de déclaration des maladies infectieuses

$\Rightarrow$ *will* $\equiv_s$ *provide* || *prévoira*

From (10) and (11) we get : *will* $\equiv_s$ *provide* $\equiv_s$ *for*

Finally we have: {*will, provide, for*} || {*prévoira*}

Consider another example:

(12) Eng.: […] members of our police (**forces** / academy) […]

Fr.: […] Les membres de nos (**services** / écoles) de police […]

$\Rightarrow$ *forces* || *services*

(13) Eng.: […] members of our (**police forces** / surveillance personnel) […]

Fr.: […] Les membres de nos services de (**police** / surveillance) […]

$\Rightarrow$ *police* $\equiv_s$ *forces* || *police*

(14) Eng.: […] members of our (**police forces** / secret services) […]

Fr.: […] Les membres de nos services (**de police** / secrets) […]

$\Rightarrow$ *police* $\equiv_s$ *forces* || *de* $\equiv_t$ *police*

From (12), (13) and (14) we get : *police* $\equiv_s$ *forces* and *services* $\equiv_t$ *de* $\equiv_t$ *police*

Thus we have: {*police, forces*} || {*service, de, police*}

Finally, the resulting units agree exactly with the translational compositionality criterion. They correspond to what Vinay & Darbelnet (1958 : 37) or Sager (1994 : 212) called “translation unit”. As Vinay & Darbelnet (1958 : 37) wrote: “We could say that translation unit is the

smaller segment whose internal cohesion prevents from a separate translation of its constituents" (we translate[2]). Moreover, it opens out onto an interesting method to identify compound words from a bilingual point of view.

2.4 Limits of the commutation test

However, Mahimon (1999 : 53) noticed that the commutation test meets with difficulties, mainly due to syntactic causes or cases of "free translation".

For instance, in the following sentences, the prepositions *d'* and *to* cannot switch alone, because of their inclusion in a wider syntactic structure:

(15) Eng.: The Petitioners are asking **to** establish (…)
Fr. : Les pétitionnaires demandent au parlement **d'**établir (…)

These limitations are mainly due to the nature of the test: since we are looking for translation units, we should switch only translation units, without affecting the syntactic relations between these units and the rest of the sentence. Semantic variations yielded by commutation should only depend on the content of the switched units, and not on external changes indirectly induced. The interpretation of the switched units context must remain identical, at both syntactic and semantic levels.

In the previous example, we could do the following commutations :

(16) Eng.: The Petitioners are (**asking** / coming) to establish (…)
Fr.: Les pétitionnaires (**demandent /** vont) au parlement (**d' /** pour) établir (…)

That would imply, by transitivity, the relation : *asking* || *demandent* $\equiv_t$ *d'*

In spite of appearances, these commutations are not correct. The preposition *au* and *d'* are linked with the predicative structure of the verb *demander* (*Y demande à X de faire Z,* lit-

erally Y *asks X to do Y*). When *demandent* is switched with *vont*, the interpretation of the preposition *au* changes and the prepositional phrase *pour établir* inherits a different grammatical function.

As to semantic relations, commutation should not affect the interpretation of surrounding units when they are polysemous or ambiguous. Consider the following sentences :

(17) Fr.: […] la base (**bruxelloise** / de données) du mouvement qui mène des campagnes […]
Eng.: […] (**its Brussels centre** / the database of the movement) which runs campaigns […]

If we take other normal commutations of *base* and *mouvement* into account, we should conclude that *Brussels centre* is aligned with *base bruxelloise du mouvement* as a whole. But of course, commutation of example (17) is not licit, because *base de données* is a compound, and the switching of *bruxelloise* with *de données* modify the interpretation of the head of the noun phrase, *base*. A free combination should not be replaced by a frozen multi-word unit.

We have seen that some commutations are illicit, when external semantic and syntactic relations are altered. Some other commutations are licit, but insufficient to recognize a real compound. In example (18), *security* and *service* are switched separately:

(18) […] members of our (**security** / intelligence) services […]
[…] Les membres de nos services de (**sécurité** / renseignement) […]

[…] members of our security (**services** / units) […]
[…] Les membres de nos (**services** / unités) de sécurité […]

Since we have *security* || *sécurité* and *services* || *services* we conclude that *security* is aligned with *sécurité* and *services* with *services*.

But test (19) results in other units :

(19) […] members of our (**security services** / maintenance department) […]
[…] Les membres de nos services de (**sécurité** / entretien) […]

We get :

$$security \equiv_s services \parallel services \equiv_t de \equiv_t sécurité$$

In this case, *security services* as a whole appears to be a compound translation unit.

These examples show that commutation is not a neutral operation: the test conclusion closely depends on the choice of the units that are involved in the switching. We face again the problem of segmentation: commutation cannot really be a criterion to find the border of translation units, because we have to determine *before* switching which are multi-word units and which are not. Without this knowledge, we could break artificially compound words and expressions, or introduce new ones.

Moreover, commutation test becomes impracticable when source and target sentences present diverging constructions. Consider example (20), where *marking* is switched, involving important changes in both source and target text:

(20) Eng.: (…) the marking of banknote for the benefit of the blind and partially sighted
Fr. : (…) l'émission de billets de banque identifiables par les aveugles et par les personnes à vision réduite

Eng.: (…) the (**marking** / destruction) of banknote (**for the benefit of /** that are identifiable by) the blind and partially sighted
Fr. : (…) (**l'émission** / la destruction) de(**/s**) billets de banque identifiables par les aveugles et par les personnes à vision réduite

Eng.: (…) the (**marking** / issue) of banknote (**for the benefit of** / useless for) the blind and partially sighted
Fr. : (…) l'émission de billets de banque (**identifiables /** inutilisables par) les aveugles et par les personnes à vision réduite

Finally we have : *marking* $\equiv_s$ *for* $\equiv_s$ *the* $\equiv_s$ *benefit* $\equiv_s$ *of* || *l'*$\equiv_t$ *émission* $\equiv_t$ *identifiables*

What does this pair mean, outside of its particular context? The problem lies in the nature of translational equivalence, which does not apply here to lexical units, but to the global meaning of the sentence. It is fallacious to switch units inside two sentences that do not construct their meaning in the same way.

2.5 The concept of lexical correspondence

Finally, lexical alignment remains a problematic task. According to our definition, the alignment concept is based on translational compositionality. But the implementation of parallel commutation test, which aims at bringing to the fore this compositionality, allows a wide range of subjective interpretation and is still arbitrary. In the Blinker Project (Melamed 1998b) five human annotators were asked to manually align 250 verses of the Bible. Each annotator had been given a complete aligning guide, with detailed criteria to solve aligning problems, and a specific software. The average agreement rate between annotators, taken two by two, was around 82%, showing the inherent subjectivity of such a task.

In order to cope with this difficulty, we suggest to withdraw the problem of compositionality, and to distinguish the monolingual identification of lexical units from the bilingual pairing of units that are translational equivalents. Given *u*, a single or multi-word unit in the source text, it is possible to look for a potential equivalent in the target. Then, if there is no satisfying match, it is not necessary to redefine *u* in a larger cluster, because the *one-to-one matching assumption* does not stand anymore.

From this point of view, the relevant questions become:

- What kind of unit do we select in each language? According to the needs, it can be limited to terms, content words, noun phrases, phraseology, etc.

- What kind of equivalence is requested, and at which degree? Different criteria are conceivable: the semantic identity or similarity, the possibility to reuse the pair in a different context, the nature of conceptual links (hypernymy, meronymy, etc.), and so on.

The lexical correspondences extraction could seem very similar to the "translation spotting" task defined in the context of the Arcade Project, where a complete mapping between the units of the two sentences is not requested. In translation spotting, the research of corresponding target units is limited to a set of previously selected source units, so there is no one-to-one assumption. But the units can be extended in order to satisfy the compositionality criterion. For instance, consider the next example:

(21) Fr.: absence de financement approprié pour les étudiants qui ont de petits moyens et impossibilité de transférer les bourses et prêts d'un pays à l'autre de la Communauté;
Eng.: lack of adequate finance for less well off students and no transferability of grants/loans throughout the Community

To spot the English translation of “petits”, the Arcade tagging guidelines propose a “phrasal correspondence”: (qui ont de petits moyens ; less well off). But according to our definition, the lexical correspondence of “petit” is just empty. If we look for the correspondence of “avoir de petits moyens” taken as an idiomatic expression, we find the following lexical correspondence: (ont de petits moyens ; less well off). The grammatical divergences (verbal *vs* adjectival phrase) do not matter, because there is no need to look for the conservation of the part of speech. Lexical material involved in translation can have diverging grammatical nature. Thus, it is not necessary to include grammatical structures that only aims at satisfying the compositionality criterion, such as the relative pronoun “qui” in the example above.

Arcade guidelines rules (e.g. “when an English participle is translated by a relative pronoun followed by a verb in French, the relative pronoun should be included”) show that translation spotting has to be conform to the commutation test. But lexical correspondences extraction has not.

3. Automatic extraction of lexical correspondences

After this theoretical discussion, we can address the problems arising with the automation of the lexical correspondences extraction.

The task having been redefined, we are now able to construct manually a set of lexical correspondences, essential to implement an evaluation of automatic techniques. First, detailed criteria have to be given for unit identification. For our experiments, we chose to identify multi-word units as a whole whenever these units could potentially be of some interest for a translator. We manually identified multi-words units independently in each language, following semantic and syntactic criteria. Then they were clustered in single units. We focused on multi-word classes for which the translation is not always word-for-word:

- Frozen phrases: *chemins de fer, in order to*

- Verbs with preposition (when it is not a free combination): *to result in*

- Collocation: *to add its support*

- Phraseology (expressions reflecting linguistic habits): *la question se pose, of little assistance*

- Terms: *Community Support Frameworks, assistance routière*

Out of these cases, every single word was considered as a single lexical unit.

Then, to pair the units with each other, we followed a simple criterion: the translational equivalence had to be valid at a general level, independently of the particular context of our corpus. When a lexical unit did not have a satisfactory equivalent among the corresponding sentence, we just put it aside: about 20% of the units were withdrawn.

The test corpus is composed of a sample of 770 pairs of aligned sentences drawn from the French and English versions of the JOC corpus used in the Arcade Project. It is a record of written questions asked by members of the European Parliament, with the corresponding answers of the European Commission. These questions, published in 1993 in one section of the C Series of the Official Journal of the European Community, have been recorded within the MLCC-MULTEXT projects. They concern various matters regarding environment, economic policy, transport, agriculture, human rights, foreign policy, institutions, etc..

Given the *gold standard*, i.e. the manually constructed set of correspondences that are considered to be exact, we can implement proper *metrics* in order to compare quantitatively any other set of lexical pairs with the standard.

The metrics used for this comparison are the classical measures of precision, recall and F-measure.

$$P = \frac{|C \cap C_{ref}|}{|C|} \quad R = \frac{|C \cap C_{ref}|}{|C_{ref}|} \text{ and } F = \frac{2 \times (P \times R)}{(P + R)}$$

where C represents the set of the evaluated correspondences, and C_{ref} the set of correspondences of the gold standard.

Note that the manually aligned corpus is not a training corpus: it is just a test corpus that allows a precise evaluation of the results.

3.1 Similarity measures

When two units are translational equivalents, they probably have similar distributions through the parallel corpus. It is possible to evaluate this similarity by counting the co-occurrences of both units (i.e. their occurrences at the same time in aligned sentences), related to the respective numbers of times they occur separately.

We tested different measures to compute this similarity:

- MI: the mutual information which quantifies the amount of information brought by an event on another event (Shannon, 1949).

- TS: the t-score, designed to filter out insignificant mutual information values (Fung & Church 1994: 1098).

- LR: the log-likelihood ratio (Dunning 1993: 69), based on a binomial distribution model, more adapted for rare events.

- P0: the log-probability of the null hypothesis, i.e. the probability for two units (u_1,u_2) to co-occur only by chance. We computed this probability assuming a binomial distribution. Without simplification, this probability can be expressed by the following equation:

$$P_0(n_{12} / n, n_1, n_2) = \frac{\binom{n}{n_1} \cdot \binom{n_1}{n_{12}} \cdot \binom{n-n_1}{n_2 - n_{12}}}{\binom{n}{n_1} \cdot \binom{n}{n_2}}$$

where n is the number of sentence pairs, n_1 and n_2 are the respective numbers of occurrences of u_1 and u_2, and n_{12} is the number of times that u_1 and u_2 co-occur in the same sentence pairs. This probability is computed as the result of 3 independent draws, assuming that each unit occurs only once in the same sentence pair :

$\binom{n}{n_1}$ is the number of different possible draws for the n_1 occurrences of u_1.

$\binom{n_1}{n_{12}}$ is the number of different possible draws for the n_{12} occurrences of u_2 that co-occur with u_1.

$\binom{n-n_1}{n_2 - n_{12}}$ is the number of different possible draws for the n_2-n_{12} occurrences of u_2 that do not co-occur with u_1.

The denominator $\binom{n}{n_1} \cdot \binom{n}{n_2}$ is the total number of possible draws without making any assumption on n_{12}.

- CO: the log probability of *cognateness*, similar to the measure proposed by Simard et al. (1992: 70), i.e. the probability to observe superficial resemblance between two compared strings, under null hypothesis. The event of cognateness is determined by counting the length of the common maximum sub-string, using techniques that we have previously developed for sentence aligning (Kraif 1999). Two units are considered as potential cognates if the sub-string exceeds a certain proportion of the smallest unit. For instance, between *contrôle* (French) and *control* (English), there is a sub-string of length 6 : c-o-n-t-r-l, which represents 6/7 of *control*.

We tested two different thresholds for this proportion: 2/3 and 1/2. Thus, we obtain two versions of CO, COa and COb, yielding different tunings between noise and silence in the identification of cognateness: COa, for which the threshold is 2/3, is less noisy and more silent than COb.

The probability of cognateness between two randomly drawn units has been computed from empirical observations (on another corpus).

- PC = P0 + CO: this metric cumulates two different kinds of information, co-occurrences and resemblance, assuming that they are independent. Given two units that co-occur n_{12} times and that are potential cognates, it estimates the unlikelihood that this event could happen only by chance.

The statistics of co-occurrence were computed on the whole French and English versions of the JOC corpus, including 69,160 automatically aligned sentence pairs (according to the methods described in Kraif 1999).

3.2 Algorithm

All these statistics have been implemented in a simple algorithm. If we consider a given source sentence and the corresponding target:

1. To create a set of candidate pairs, every unit of the source sentence is compared with every unit of the target, giving for each pair an association score. The scores are then ranked in descending order.

2. The best scoring pair (u_1,u_2) is recorded.

3. All the other candidate pairs that involve either u_1 or u_2 are removed.

Step 2 and 3 are reiterated until there is no more candidate pair.

Two source units that co-occur frequently on the syntagmatic axis will tend to be associated to the same target units. In order to reduce the effect of these indirect associations, step 3 implements a kind of competition between the potential pairs, allowing each source unit to be associated with only one unit in the target sentence, and vice-versa. As demonstrated by Melamed (1998a: 14), this algorithm approximately establishes the best scoring set of correspondences under the competitive linking criterion.

To increase the performance of the algorithm, we made the following approximations:

- In the same pair of aligned sentences, we took into account only one occurrence of each lexical unit. In our test corpus, 9% of the pairs included repetitive units, and were thus ignored.

- Very frequent units, which had more than 5,000 occurrences in the JOC corpus, were withdrawn: 29 English units and 38 French units were concerned (mostly punctuation marks and frequent function words).

As a result of the withdrawal of these units, 31% of the correct pairs have not been considered. Then, the recall of every extraction was, in any case, below 69%.

3.3 Results

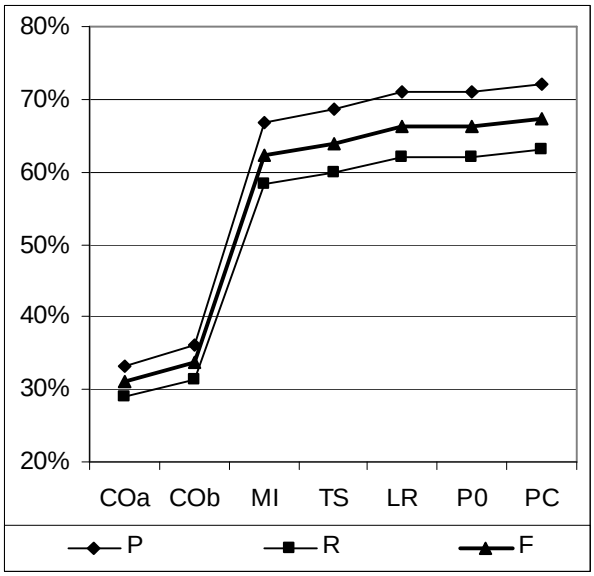

Figure 1

Results are depicted on figure 1. In such a task, P and R are strongly linked. We can rank the measures in ascending order as follows: COa, COb, MI, TS, P0, LR and PC. P0 and LR have a very close behaviour: their distributions are asymptotically the same.

PC got the best results with P = 72,2%, R = 63% and F = 67,3%. The combination of CO and P0 improves slightly the results: that indicates that cognateness and distribution complement each other. For CO alone, we notice that COb is more efficient than COa: the extra noise brought by COb seems to be filtered out by the competition between different pairings. Finally, when we compute the co-occurrences vectors on lemmatised unit, the global results increase slightly by 1%.

3.4 Filtering of the results

A filtering method has to fulfil two conditions: eliminating the most erroneous pairing while keeping the most correct pairs. For this task, we can use the calculated scores as a good indicator of the reliability of an association. On the basis of the competitive linking criterion, we developed a “differential” filtering method: we can suppose that if different target units compete with each other to be associated with a same source unit, there is a greater uncertainty about the association. Thus, for each recorded pair, we compute the ratio between its score and the score of the second best competing pair. If the ratio is lower than a certain threshold they are both eliminated. We tested 8 values for the threshold : 1 (no filtering), 1.05, 1.2, 1.5, 2.5, 3 and 4. Figure 2 shows the concomitant evolutions of precision and recall for these different thresholds.

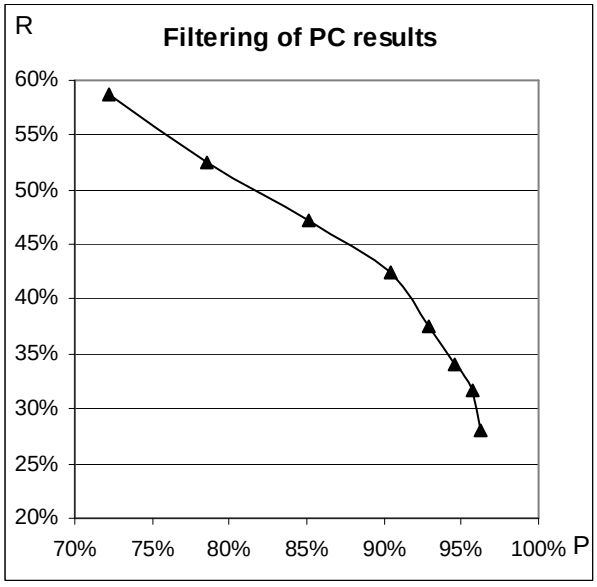

Figure 2

This method clearly shows that it is possible to increase precision to very high levels by sacrificing recall: for instance, with PC, we can get a 96% precision with a recall of about 35%.

For the cognate-based measure, the differential filtering allows a 90% precision for a 25% recall, demonstrating that the important noise brought by n-gram comparison can easily be reduced by a simple algorithmic framework.

3.6 Effect of corpus size

Finally, we would like to determine the effect of the most important parameter for these statistical tools: the size of the aligned corpus where occurrences and co-occurrences are observed. We tested 7 different sizes: from the sole test corpus, comprising 770 sentence pairs, to the complete JOC corpus, comprising 69,160 pairs. Results are displayed on figure 3.

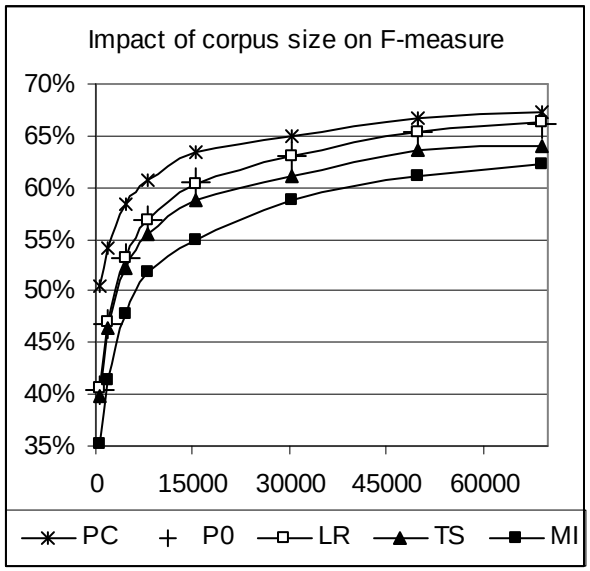

Figure 3

We note an important progression for every measure. The highest increase is for MI, from 35.2% to 62.2%. P0 and LR, initially very close to TS, tend to progress faster.

For the smallest corpus, the difference between PC and the other measures is important: about 10% better. But this interval gradually decreases while the corpus becomes greater. It was predictable: the cognate-based information is a constant which does not depend on the

size of corpus. The more efficient distribution-based statistics are, the less additional information cognates bring.

The progression begins fast and ends very slowly: for instance, LR increases by 22,5% from 770 to 30,238 sentence pairs, but only by 3,2% from 30,238 to 69,160 pairs. With the help of cognates, the best results are almost reached by computing the statistics on hardly half the training corpus, involving a serious saving of computation time and space.

3.4. Conditional entropy of a set of correspondences

If we compare the gold standard with a set of randomly drawn correspondences, we notice some differences at a formal level. As expected, the correspondences are far more regular in the case of the gold standard: a source lexical unit is often paired with the same target units. Of course, in this case, paired units are strongly linked by a same semantic content. When units are randomly paired, without any constraint, the correspondences are unsystematic. For instance, as to the 10 occurrences of *against* in the gold standard, we count only 3 different French translations, whereas in a random set of correspondences we get 10 different associated units. Thus, the gold standard contains probably more “order” than any erroneous set of correspondences.

Manually extracted correspondences	*Randomly extracted correspondences*
(against, à l'encontre de)	(against, par)
(against, à l'encontre de)	(against, procédure)
(against, à l'encontre de)	(against, moratoire)
(against, au détriment de)	(against, à l’encontre de)
(against, contre)	(against, dont)
(against, contre)	(against, contre)
(against, contre)	(against, effectivement)
(against, contre)	(against, charges)
(against, contre)	(against, Etat membre)
(against, contre)	(against, qui)

Table 1: manually extracted correspondence set presents less entropy

This indicates another kind of evaluation, based on the following hypothesis: the more regular a set of correspondences is, the closer to the gold standard it should be. To quantify the regularity of a set of pairs, we propose to calculate the conditional entropy of the two distributions of lexical units :

$$H(F/E) = -\sum_{e} p(e) \sum_{f} p(f/e) \log p(f/e) = -\sum_{e} \sum_{f} p(e,f) \log \frac{p(e,f)}{p(e)} \qquad (1)$$

$$H(E/F) = -\sum_{f} p(f) \sum_{e} p(e/f) \log p(e/f) = -\sum_{f} \sum_{e} p(e,f) \log \frac{p(e,f)}{p(f)} \qquad (2)$$

where *e* and *f* are referring to lexical units of the English and French texts.

To observe the possible correlation between conditional entropy and the correctness of a correspondence extraction, we need to get different sets of correspondences, with various values for precision and recall. Using the previous algorithm (called Algo 2), we developed a measure combining PC and a random draw, in different proportions : we obtained seven sets with F-measure ranging from 6% to 65%.

In order to evaluate a wider range of pairings, we implemented several other extractions using CO, IM, TS, LR, P0 and PC with another simpler algorithm (called Algo 1), where each source unit is paired with the best-scoring target unit. The results of this algorithm are inferior and have different formal characteristics: the pairing between the units of two aligned sentences are not one-to-one, but sometimes many-to-one because of strong indirect associations.

Then, we filtered the results of Algo 1 and Algo 2 (using differential filtering). We finally obtained 31 sets of correspondences. For each of these sets, we computed H(e/f) and H(f/e).

As shown in figure 4, we observe a strong correlation between the precision P and the value of max(H(e/f),H(f/e)). The linear correlation coefficient between P and max(H(e/f),H(f/e)) is about -0,95.

Notice that recall (as well as F) can be deduced from precision, taking into account the number of proposed pairs, but it is not *directly* linked to the conditional entropy.

We plotted a dot for the gold standard, for which the conditional entropy is low but not minimal. This is due to the normal variations induced by the process of translation. If some extractions yield lower entropy, it can be explained by a very low recall.

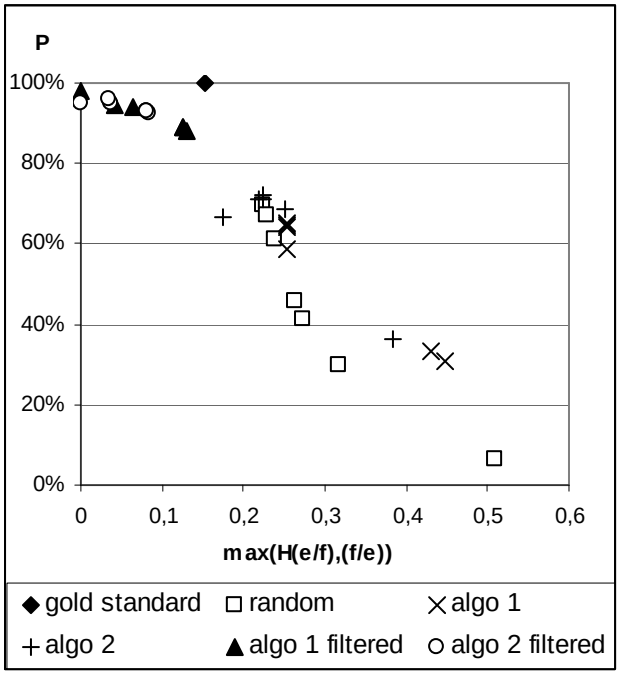

Figure 4
Correlation between Conditional entropy and Precision

In this way, the conditional entropy constitutes a good indicator for a comparative evaluation of different sets of lexical correspondences, without appealing to a manually extracted set.

4. From translational data to contrastive knowledge: illustration

We have pointed out that human translation, in general, could not be reduced to a simple transformation from one language to another language. A translation is the result of particular choices of the translator, driven by a particular communicative background. As Seleskovitch said, translating is more than just “transcoding” (quoted by Laplace, 1994: 240) .

However, the previous results in lexical correspondences extraction show that contrastive knowledge (i.e. linguistic knowledge about the different ways used by different languages to denote similar semantic contents), can be automatically extracted from translational data. Moreover, as showed by the measure of entropy, this contrastive knowledge emerges from objective phenomena, and does not depend on a subjective understanding.

Our experiments showed contrastive properties about lexicon, but the same kind of statistical filters could be used to observe and study any contrastive phenomena, concerning various linguistic features: tenses, parts of speech, concord of tense, aspects, diathesis, word order, semantic features, etc. To implement that kind of study we just need a properly tagged aligned corpus (which is not an easy thing to find!).

4.2 Example of contrastive phenomena

To give an example of such contrastive observations, we tagged[1] the parts of speech in our 770 test sentences. Then, it was very easy to give the detailed results of precision and recall of the lexical pairing, for each identified class, in both directions. These results are displayed in table 2:

Precision				*Recall*			
English		*French*		*English*		*French*	
Class	*P*	*Class*	*P*	*Class*	*R*	*Class*	*R*
stop word	46,3%	stop word	47,0%	stop word	70,3%	verb	73,2%
adverb	62,2%	verb	68,9%	adverb	70,9%	stop word	79,9%
verb	70,7%	adjective	77,5%	verb	77,1%	adjective	80,9%
adjective	80,1%	adverb	79,0%	adjective	85,1%	noun	84,5%
noun	85,0%	noun	83,9%	noun	86,6%	adverb	87,5%
proper noun	89,7%	proper noun	91,7%	proper noun	90,5%	proper noun	91,3%

Table 2: results by part of speech (for P0)

Named entities, toponyms, ethnonyms and proper nouns were put in a separate class called “proper noun”. Conjunctions, prepositions, articles and other function words were arranged under “stop word”. For both languages, we can roughly order the results in the following way (without adverbs):

stop words < verb < adjective < noun < proper noun

As these results are only based on a distributional criterion (P0), they can be linked with entropy: the more variable the translation of a lexical unit is, the worse the results are. Thus we can easily explain the differences between different classes of words: proper nouns generally have stable translations, while stop words are more inconsistent. We propose the same interpretation for verb, adjective and noun, which present intermediate degrees of variation. The adverbs do not show a very clear behaviour, so we cannot conclude about the stability of their translations.

From a contrastive point of view, it could be interesting to study the correspondence between parts of speech from one language to another. Table 3 displays, for each class and both directions, the rate of units that are translated into the same part of speech.

[1] We did it in a very simple, and approximate, manner: given the alphabetic list of every unit, we manually indicated the part of speech, outside any context. Ambiguous cases were not taken into account in the following results.

	noun	*proper noun*	*verb*	*adverb*	*adjective*	*stop word*
French to English	75.12%	96.50%	68.46%	55.17%	40.82%	98.30%
English to French	87.51%	81.42%	76.26%	40.51%	63.15%	87.08%

Table 3 : rates of stability for each part of speech

These statistics reveal contrastive phenomena, that we can examine more precisely. The following examples illustrate the more significant transformations.

From English to French we observe that :

- 8 % English adverbs are paired with ambiguous adjective / noun forms:

  Eng.: (…) to create a new framework to facilitate, both legally and **financially**, the distribution (…)

  Fr.: (…) créer un nouveau cadre visant à faciliter, sur les plans législatif et **financier**, la circulation (…)

- 7 % English verbs are paired with a noun:

  Eng.: Thus the United States **applies** the reduced rate (…)

  Fr.: Les États-Unis d'Amérique accordent ainsi directement l'**application** du taux réduit (…)

  Eng.: (…) to prevent the Athens-Delphi road being **widened**

  Fr.: (…) afin qu'il ne soit pas procédé à l'**élargissement** de la route Athènes-Delphes

- 6 % English adverbs are paired with an adjective:

  Eng.: (…) thus excluding the only properly **democratically** elected institutions.

  Fr.: (…) d'où est donc exclue la seule institution issue d'élections **démocratiques** appropriées.

- 4% English adjectives are paired with a noun:

  Eng.: (…) assurances that the Bishops would be in no danger and **free** to move about (…)

  Fr.: (…) l'assurance que les évêques ne seraient pas en danger, qu'ils bénéficieraient de la **liberté** de mouvement (…)

From French to English we note that :

- 17 % French adverbs are paired with an adjective:

  Fr.: La Commission a-t-elle l'intention d'adopter des mesures destinées à venir **financièrement** en aide aux agriculteurs (…)

  Eng.: Does the Commission intend to provide **economic** assistance for those farmers (…)

- 10 % French adjectives are paired with a noun:

  Fr.: (…) certificats **sanitaires** croates.

  Eng.: (…) Croatian **health** certificates.

In the last case, it appears that almost every French adjective is a relational adjective: alimentaire, artisanal, auditif, budgétaire, céréalier, climatique, communautaire, législatif, maritime, sanitaire, tarifaire, touristique, écologique, minoritaire, etc.

Of course, the pointed transformations would require a finer linguistic analysis. The statistical tools and filtering methods that we have presented just aim at bringing to the fore raw bilingual material rich in contrastive phenomena.

5. Conclusion

Our first goal was to align parallel texts at the lexical level. An accurate analysis of a human translation corpus showed us that it was difficult to determine such an alignment on the basis of translational compositionality. Indeed, translation is not a transformation based on lexical units, even if we take into account deep syntactic transformations: human translation often involves a complete reconstruction of the global meaning of a given textual segment, where particular choices are made according to a particular situation of communication. To avoid an inconsistent segmentation linked with semantic discrepancy problems, we suggested to implement another task, the *lexical correspondence extraction*, where the units are determined upstream, according to the needs. After having manually extracted such a set of correspondences, in a test corpus, we have implemented simple techniques that allow to obtain sur-

prisingly good results automatically. Using association scores as log-likelihood ratio or log-probability of null hypothesis, combined with cognateness, in the framework of a competitive linking algorithm, we reached a F-measure around 67,3% (in a simplified implementation where recall could not exceed 69%).

These results seem to contradict our assertion that translating is not transcoding, because all these correspondences are valid at the linguistic level, outside of the message specificities. But we do not think it is a real contradiction: the source text *determines* the target but this relation is not *deterministic*. There are *regularities* in the transformation from the source to the target text; but there are no transformation *rules*, as in Machine Translation. These results only prove that it is possible to filter out the noise brought by contextual and specific choices, in order to point out these regularities through the mass of particular translations.

Interestingly, a strong correlation between the precision of the results and conditional entropy has been observed. Thus, translational regularities can be picked up, and extracted, in an objective way. Given that the co-occurrence/occurrence counting is a very general principle, it can be extended to any contrastive feature. Studying the correlation between parts of speech, we gave a very poor and simplified illustration of this kind of observation. But it is possible to focus on any linguistic property, in order to compare it through two or more different languages. We fully agree with Isabelle (1992: 8) when he said: “Given the staggering volume of translations produced year after year, it is quite obvious that *existing translations contain more solutions to more translation problems* than any other existing resource”. This mass of translational data requires the development of specific tools to be explored: simple statistical measures already open the way up to this exploration. Of course, these techniques need to be refined, but the next important step is to build large collection of annotated multi-

texts. Then, it will be possible to take advantage of the wide variety of contrastive phenomena that lies behind translation corpora.

Notes

[1] The « translation spotting » task was a kind of lexical aligning: the competing systems had to align the 3 722 occurrences of 60 polysemic units (20 adjectives, 20 nouns and 20 verbs) with their translations, through the JOC (for *Journal Officiel de la Communauté*) corpus. The best results were around 77 % of precision and 73 % of recall.

[2] "On pourrait encore dire que l'unité de traduction est le plus petit segment de l'énoncé dont la cohésion des signes est telle qu'ils ne doivent pas être traduits séparément."

Acknowledgements

The author would like to thank Jean Véronis for his support during the second campaign of the Arcade Project and his help for the access to various parallel corpora. Many thanks to Carole Guéret, for her reviewing of the first version of this text, and her helpful linguistic advice.

References

Boutsis, S. and S. Piperidis. 1996. "Automatic extraction of lexical equivalences from parallel corpora." *Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC'96), 12th European Conference on Artificial Intelligence (ECAI'96), 11-16 August 1996*, 27-31. Budapest.

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

Catford, J. C. (1965) *A Linguistic Theory of Translation*, London: Oxford University Press.

Dagan, I., K.W. Church and W. Gale. 1993. "Robust Bilingual Word Alignment for Machine Aided Translation." *Proceedings of the Workshop on Very Large Corpora*, Academic and Industrial Perspectives. 1-8. Colombus, OH.

Debili, F. 1997. "L'appariement : quels problèmes ?" *1ères JST 1997 FRANCIL de l'AUPELF-UREF, Avignon, 15-16 avril 1997*. 199-206. Avignon.

Dunning, T. 1993. "Accurate Methods for the Statistics of surprise and Coincidence." *Computational Linguistics*. 19(1): 61–74.

Fung, P. and K.W. Church. 1994. "K-vec : A New Approach for Aligning Parallel Texts." *Proceedings of the 15th International Conference on Computational Linguistics*, ICCL, 1096-1102. Kyoto.

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

Israël, F. and M. Lederer J. (eds.) 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.

Isabelle, P. 1992, "Bi-Textual Aids for Translators", *Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Waterloo, Canada (available at http://www-rali.iro.umontreal.ca/Publications.fr.html).

Isabelle, P. and M. Simard. 1996. "Propositions pour la représentation et l'évaluation des alignements de textes parallèles dans l'ARC A2." *Rapport technique*. Laval, Canada : CITI. (available at : http://www-rali.iro.umontreal.ca/arc-a2/PropEval).

Kay, M. 2000. Preface. *Parallel Text Processing*, ed. by J. Véronis. xi-xv. Dordrecht, Nederlands: Kluwer Academic Publishers.

Kraif, O. 1999. "Identification des cognats et alignement bi-textuel : une étude empirique." *Actes de TALN'99*, 205-214. Cargèse, France.

Kraif, O. 2000. "Evaluation of statistical tools for automatic extraction of lexical correspondences between parallel texts." *Proceedings of MT 2000, 20-22 november 2000*, 161-168. Exeter UK.

Kraif, O. 2001. "Translation alignment and lexical correspondences : a methodological reflection." *Lexis in contrast. Studies in Corpus Linguistics*, ed. by B. Altenberg and S. Granger. Amsterdam: John Benjamins.

Laplace, C. 1994. *Théorie du langage et théorie de la traduction*. Paris, Didier érudition.

Mahimon, M.-D. 1999. *Identification des équivalences traductionnelles sur un corpus Français / Anglais, Mémoire de DEA*. Aix-en-Provence : Université de Provence Aix-Marseille 1.

Malavazos, C., S. Piperidis, G. Carayannis, G. 2000. “Towards memory and template-based translation synthesis.” *Proceedings of MT 2000, 20-22 november 2000*, 1.1-1.8. Exeter UK.

Melamed, I. D. 1998a. *Technical Report#98-08*, Philadelphia: Institute for Research in Cognitive Science University of Pennsylvania. (available at URL: http://www.cs.nyu.edu/~melamed/).

Melamed, I. D. 1998b. *Manual Annotation of Translational Equivalence: The Blinker Project. Technical Report # 98-07*, Philadelphia: Institute for Research in Cognitive Science University of Pennsylvania. (available at URL: http://www.cs.nyu.edu/~melamed/).

Nida, E. 1969. *The theory and practice of translation*, Leiden: Brill.

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

Shannon, C. E. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

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

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

Simard, M. .2000. “Multilingual text alignment – Aligning three or more versions of a text.” *Parallel Text Processing*, ed. by J. Véronis, 49-68. Dordrecht, Nederlands: Kluwer Academic Publishers.

Véronis, J. and P. Langlais. 2000. Evaluation of parallel text alignment systems – The ARCADE project. *Parallel Text Processing*, ed. by J. Véronis, 49-68. Dordrecht, Nederlands: Kluwer Academic Publishers.

Vinay, J.-P, J. Darbelnet. 1958. *Stylistique comparée du français et de l’anglais*. Paris : Didier.