Annotation of anaphoric expressions in an aligned bilingual corpus

Agnès Tutin*, Meriam Haddara*, Ruslan Mitkov+, Constantin Orasan+

* LIDILEM, Université Grenoble 3-Stendhal BP 25
F-38040 Grenoble Cédex 9

+Computational Linguistic Group, School of Languages and European Studies
Stafford Street, Wolverhampton WV1 15B, GB

agnes.tutin@u-grenoble3.fr;meriam.haddara@laposte.net;R.Mitkov@wlv.ac.uk;C.Orasan@wlv.ac.uk

Abstract

This paper discusses a French-English corpus annotated and aligned at anaphoric level. It also presents an annotation scheme based on the study of a detailed corpus featuring different types of correspondences and mismatches. The scheme which is adapted from EAGLES recommendations, supports the alignment at anaphoric level and caters for the different kinds of mismatches.

1. Introduction

In recent years we have seen the emergence of corpora annotated for anaphoric or coreferential links which have proved to be very useful for linguistic studies and a variety of NLP applications alike. Whereas several projects have already delivered such corpora (Ge 1998, Mitkov *et al*. 2000, Popescu-Belis 1998, Tutin *et al.* 2000), the availability of these data is still scarce.

In addition to the fact that some of the above corpora are difficult to obtain[1], all of them share the same limitation: they are monolingual and such, cannot be exploited in multilingual applications. Against the background of already existing multilingual aligned corpora at sentence level (Erjavec et al. 1998, Simard 1998), we believe that it is very important to develop bi- and multilingual resources aligned at anaphoric level. The benefit of aligned multilingual corpora annotated at anaphoric or coreferential level can be seen in many ways. To start with, an aligned bilingual corpus can be used for pedagogical purposes as a basis for data-driven learning. Grammatical anaphoric expressions greatly differ depending on the languages: for example, clitic pronouns are specific to romance languages and can be easily learned in an inductive way by comparison with another kind of language. As across languages translations of anaphoric pronouns are not necessarily rendered as overt target pronouns (Mitkov 1996, Uehara 1996), such a corpus would be of interest to translation studies as well. Next the availability of such a corpus would be beneficial to NLP with anaphora resolution and machine translation being obvious examples. In machine translation a bilingual corpus marked up for anaphor-antecedents pairs could be used both for developing rules as in the case of example-based translation, or generally for evaluating purposes. In anaphora resolution, such corpus could be used both for training bilingual systems as reported in (Harabagiu and Maiorano 2000) and for evaluating them. It can be exploited even in the evaluation of monolingual systems for comparative purposes – for instance, it becomes possible to compare anaphora resolution systems for English with those for French on almost identical texts. Finally, such an aligned corpus can be employed by a bilingual anaphora resolution system to enhance the performance for each language by taking into account language-specific gender discrimination and/or translations of pronouns as proposed and shown by Mitkov and Barbu (2000). It is clear that if the corpus is trilingual, it could be of even greater benefit for such a strategy, especially in the case of a language with a typical three-gender system as German or Slavonic languages.

As part of a collaborative project, the Natural Language Processing group at LIDILEM, University of Stendhal, Grenoble and the Research Group in Computational Linguistics at the University of Wolverhampton, we are building a 25 000 word bilingual corpus aligned at anaphoric level, suitable as for both contrastive linguistic studies and NLP applications. The markup so far has been restricted to grammatical anaphoric expressions (mainly pronominal expressions and possessive determiners) and nominal antecedents involved in a coreferential relation which is sufficient for the majority of NLP purposes. In the second stage of the project, we plan to extend the scope of anaphoric expressions (not only overt pronouns but also zero pronouns, definite NPs, demonstratives, etc.) and the scope of relations (not only coreference, but also description, part-of, etc.).

2. Correspondences and mismatches of anaphoric expressions in the BAF corpus

The corpus we use is a part of the ARCADE corpus (Simard 1998) including a technical report in sociology, two scientific papers in NLP and a literary text (three chapters of Jules Verne's *De la Terre à la Lune*).

In its first stage the project focused on the annotation of anaphoric personal pronouns and anaphoric possessive determiners. Whereas those were annotated along with their equivalents for both English and French, at this stage we excluded anaphoric expressions which have not been translated in the other language. We then analysed the pairs obtained (530) and distinguished three types of pairs:

a) The target anaphoric expression matches either a personal pronoun or a possessive determiner.

[1] Some of them are only available on a commercial basis.

b) The target expression is replaced by a different type of anaphoric expression.
c) The target expression is translated by way of reformulation and contained no referring expression.

The two first types of pairs are cases of correspondence. The third one concerns reformulation. The results obtained for each text are summarised in the following table:

	Correspondence (a and b)		Non correspondence (c)	total
	PP or PD[2] ↔ PP or PD (a)	PP or PD ↔ other referring exp (b)	PP or DP ↔ reformulation	
TAO3	11	2	14	27
%	48,15%		51,85%	100%
TAO2	52	20	41	113
%	63,72%		36,28%	100%
CITI2	25	7	32	64
%	50,00%		50,00%	100%
VERNE	216	29	81	326
%	75,15%		24,85%	100%
Whole CORPUS	**304**	**58**	**168**	**530**
%	**68,31%**		**31,69%**	**100%**

Table 1 : Correspondences and mismatches between anaphoric expressions

The level of correspondence seems to depend highly on the textual genre (and on the kind of translation!). The literary text (VERNE) surprisingly contains more correspondences than the technical papers (it may also contain proportionally more references to human characters).

2.1 Linguistic parameters concerning elements in correspondence

In the second type of pair correspondences (case b), the anaphoric expression can be replaced in the other language by different types of anaphoric pronouns (typically demonstrative ones), grammatical ellipses of the subject (zero pronouns) or definite nominal groups, that is to say expressions that can easily be annotated. For this reason we included them in the category of correspondences as with pairs of the first type (case a). We were able to classify the different types of correspondence on the basis the following parameters:

1) The type of anaphoric expression.
There is an 'identity' of anaphoric process when a personal pronoun or a possessive determiner is translated by a comparable[3] expression in the other language. That is not the case for the following pair for example:

(…) pour deviner les instincts d'un homme, on doive ***le*** *regarder de profil (…)* → Personal pronoun
(…) in order to judge a man's character one must look at ***his*** *profile (…)* → Possessive determiner

2) The syntactic function of the anaphoric expressions.
This parameter further distinguishes the expressions according to their syntactic function, or the function of the noun phrase for possessive determiners.

*(…) Bilsby entre les quatre dents qu'****il*** *avait sauvées …de la bataille.* → Subject
(…) Bilsby between the four teeth which the war had left ***him.*** → Complement

3) The number properties.
There might be a discrepancy in number between the anaphoric expression and its antecedent, and also between the members of the pair:

le canon (…) ***ses*** *joyeuses détonations* → Plural
the guns (…) ***their*** *delightful reports* → Singular

4) The antecedent of the expressions.
This parameter checks that the antecedents of the expressions are the same. In a literal translation, the antecedents would be very similar in the two languages.

2.2 Cases of non correspondences

Pairs classified under case c regroup heterogeneous phenomena. We distinguished the following mismatches:

1) Complete reformulation:
A complete reformulation like in the following example might show a tendency of the translator 'moving away' from the original text.

*le club (…) ainsi arriva-t-****il*** *à Baltimore.*
(…) so things were managed in Baltimore.

2) Frozen expressions:
If a frozen expression containing an anaphoric expression is used in one language, there might not be a counterpart with a referring expression in the other language.

Toutes ces inventions laissèrent loin derrière ***elles*** *(…)*
These inventions (…) left far in the rear

3) Syntactic alternation.
For example a subordinate clause with a tensed verb can be replaced by a present or past participle with no overt subject.

*Transcheck (…) dès qu'****il*** *trouve une occurrence de "raw milk" non rendue par "lait cru", il signale une incohérence terminologique (…)*
Transcheck (…) and finding an occurrence of « raw milk » that is not aligned with « lait cru », it signals a potential term inconsistency (…)

More interesting is the analysis of mismatches due to several language-specific features. For example:

4) Emphasis and focalisation processes in French, which seem to trigger more anaphoric pronouns than in English:

(…) alors que les temps d'exécution, ***eux****, grimpent en flèche.*
(…) while the running time increases radically.

[2] PP:personal pronoun / PD:possessive determiner.
[3] As the pronominal systems of the two languages are different, it is more appropriate to refer to comparable expressions rather than to identical expressions.

5) French clitic pronouns such as *en* et *y*, which have no counterpart in English.

*(…) en vérifiant le segment cible aligné pour **y** détecter (…)*
(…) verifies the aligned target segment for the presence of (…)

Despite complexities due to the nature of the language itself (variability and language specificities), this study proved it possible to align the anaphoric expressions of a parallel text for at least two out of three target expressions. It also shed light on complex cases (case c) that might be problematic for anaphora resolution systems.

3. Annotation scheme for aligning bilingual anaphoric expressions

Our annotation scheme is encoded in XML. Monolingual texts are marked up at anaphoric level. Alignment is annotated separately between the anaphoric expressions. Our scheme is able to deal with several cases of mismatches.

3.1 Annotation at monolingual level

The annotation scheme chosen was based on the ELRA annotation scheme already applied on a one million word corpus (Tutin *et al.* 2000), compatible with the MUC annotation scheme, widely used in evaluation tasks (Chinchor & Hirschmann 1997). Here is an example of annotated text.

```
<S ID="1459"> <EXP CAT="DNP" FC="SUBJ" NB="SING" ID="4">The system</EXP> is called TransCheck , and <EXP CAT="PP" FC="SUBJ" NB="SING" ID="5"><REF SRC="4"/>it</EXP> is capable of detecting some of the more frequently occurring types of translation errors …
```

Each expression (<EXP>) receives a unique identifier. Anaphoric expressions include an empty XML element called <REF> which refers back to the antecedent with the help of the **REF** attribute. Complex cases such as multiple antecedents, disjunctive antecedents, can be handled easily by this annotation scheme. We think that this kind of simple annotation scheme is easier to implement than stand-off markup in XML documents like in the MATE annotation scheme (Poesio 1999) which tends to be unreadable for a human annotator without the help of some user interface, even if more elegant. The inclusion of the <ptr> element at the level of the anaphoric expression greatly facilitates the annotator's and the reader's work. Syntactic information, language dependent, has been added on referring expressions (category, syntactic function, number, gender).
The annotation process, now completed, was performed with the help of a specific tool, Palinka, developed by C. Orasan and specifically designed for discourse annotation (Orasan 2003).

3.2 Bilingual markup

Alignment is performed according to the Eagles CES recommendations that we adapted for our specific task.
Since our corpora were already annotated at anaphoric level in each language, we opted for an alignment at the level of anaphoric expressions. Aligning both at antecedent and anaphor levels would be useless in most cases, since the anaphoric chain can be systematically rebuilt. Every intended anaphoric expression (i.e. anaphoric personal pronouns and possessive determiners) is aligned, if possible, to a linguistic expression in the target language whatever its categorial and anaphoric status. As a result, some referring expressions which were not included in our monolingual annotation scheme have to be annotated for the sake of alignment.
In the standard case (there is a perfect match between anaphoric expressions), expressions are just related with the help of their identifier (for the sake of readability, we do not insert here most attributes on referring expressions).

Ce que fait <EXP ID="25">le système </EXP> , concrètement , … Si, par exemple , <EXP ID="26"²> <REF SRC="25"/>il</EXP> trouve un segment en langue source contenant …	Concretely , what <EXP ID="19">the system</EXP> does … If <EXP ID="20"><REF SRC="19"/>it</EXP> finds any SL segment containing …
<LINK XTARGETS= " 26 ; 20 " >	

One anaphoric expression can have multiple correspondences in the other language.

… il n' a pas pour fonction d' imposer des termes au <EXP ID="117">réviseur</EXP> , mais seulement de <EXP ID="118"><REF SRC="117"/>l'</EXP> aider à valider une traduction préliminaire	its function is not to impose terms on <EXP ID="113">the reviser</EXP> , but only to assist <EXP ID="114"><REF SRC="113"/>him</EXP> or <EXP ID="115"><REF SRC="113"/>her</EXP> in validating a preliminary translation .
<LINK XTARGETS= " 118; 114 115" >	

One anaphoric expression may have no corresponding referring expression.

… que l'on soit obligé de morceler <EXP ID="0">les textes volumineux</EXP> , c'est-à-dire de <EXP ID="1"><REF SRC="0"/>les</EXP> diviser	… lengthy texts have to be divided up among several translators …
<LINK XTARGETS= " 1 ; " >	

Different kinds of mismatches can be dealt with: they are annotated as attributes on the <link> element.

° A mismatch on grammatical categories is annotated with the help of a **CAT="DIFF"** attribute-value. Annotation is more complex when the corresponding referring expression was not annotated in the monolingual text, either as an antecedent or as an anaphoric expression. In this case, a specific attribute is introduced on the referring expression.

Si <EXP ID="9">TransCheck </EXP> est en mesure de les déceler , par	Transcheck, on the other hand, can detect these errors, because <EXP ID="180"

contre , c' est parce qu' <EXP ID="10"><REF SRC="9"/>il</EXP> a été conçu expressément …	TYPE=ADDED>the system</EXP> was specially designed (…)
<LINK XTARGETS= " 10 ; 180 " CAT="DIFF">	

- A mismatch of syntactic functions is also annotated with the help of a **FC="DIFF"** attribute-value.
- Discrepancy concerning the number is annotated with **NUMBER="DIFF".**
- Discrepancy concerning the antecedent (i.e. antecedents are not literal translations) is annotated with **ANTE="DIFF"**.

Les termes employés par <EXP ID="112">le locuteur</EXP>, si techniques et si justes qu'**ils** soient , ne sont pas nécessairement les seuls dont <EXP ID="114"><REF SRC="112"/>il</EXP> aurait pu se servir.	The terms employed by <EXP ID="106">the writer</EXP> , no matter how technical or exact , are not necessarily the only ones <EXP ID="108"><REF SRC="106"/>he</EXP> or <EXP ID="107"><REF SRC="106"/>she</EXP> could have used . </S>
<LINK XTARGETS= " 114; 107 108" ANTE="DIFF">	

The alignment markup has been performed manually on a part of the corpus to check the validity of the annotation scheme. Few problematic cases arose during this first stage. It is envisaged that the alignment is completed in a semi-automatic way.

4. Conclusions and further work

The work on the project so far has shown that whereas aligning anaphoric expressions is a challenging, labour-intensive and time-consuming task, it is still feasible to produce a corpus with such a bilingual markup which in turn will be an invaluable resources for both linguistic and NLP studies. To date, we have annotated a corpus of 25 000 words at the anaphoric level but the alignment was only performed on 7000 words. The alignment carried out so far has been restricted to grammatical anaphors, but in the next stage of the project, it will be extended to any coreferential expression. Annotation could be extended to other parameters – annotation of semantic type would probably be an interesting perspective – and include centering information. Larger and more varied corpora should also be annotated because we observed interesting discrepancies between textual genres that should be confirmed on more data.

Another important forthcoming stage is to introduce a semi-automatic alignment procedure, which is expected to speed up the annotation. Using parallel lexicons (e.g. WordNet for English and French) and a robust parser (e.g. XIP) we are currently developing a such a procedure which attempts to match two anaphoric expressions by trying to align their antecedents, or aligning an anaphoric expression with an NP in the other language if a corresponding anaphoric expression cannot be found. The fact that our corpus is aligned at sentence level reduces the number of candidates for alignment which need to be considered. For the aligning process a set of rules is being developed. The alignment is supervised in a specially designed tool. In addition to this we are currently extending the EAGLES alignment scheme so it fits better our requirements and allows us to mark more information about the type of alignment.

Topics due to be addressed in the near future also include the interanotator agreement which is to be measured on samples of the corpus and producing comprehensive statistics related to the different types of correspondences and mismatches outlined earlier in the paper.

References

Chinchor N., Hirschmann L. (1997), MUC-7 Coreference Task definition, Version 3.0, *Proceedings of MUC-7*. http://www.muc.saic.com.

Davies S., Poesio M., Bruneseaux F., Romary L., (1998), *Annotating Coreference in Dialogues : Proposal for a Scheme for MATE (First Draft).*

Erjavec, T., Lawson A., Romary L. (1998). *East Meet West: A Compendium of Multilingual Resources*. TELRI-MULTEXT EAST CD-ROM.

Ge, N. (1998) *Annotating the Penn Treebank with coreference information*. Internal report, Department of Computer Science, Brown University.

Haddara M. (2003) *Constitution d'un corpus bilingue aligné au niveau anaphorique : étude de faisabilité*, Mémoire de maîtrise "Industries de la langue", Université Stendhal-Grenoble 3.

Harabagiu, S. and Maiorano, S. (2000) "Multilingual Coreference Resolution". *Proceedings of ANLP-NAACL2000*, 142-149.

Mitkov, R. (1996) "Machine Translation and Anaphora". *Machine Translation Review*, No. 4.

Mitkov, R. and Barbu, C. (2000) "Improving pronoun resolution in two languages by means of bilingual corpora", *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*. Lancaster, UK. 133-137.

Orasan, C. (2003) PALinkA: a highly customisable tool for discourse annotation. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, Sapporo, Japan, 5 - 6 July. 39 – 43.

Poesio M. & Vieira R. (1998) A corpus-based investigation of definite description use. *Computational Linguistics*, **24**, 2.

Poesio, Massimo (1999) MATE Dialogue Annotation Guidelines – Coreference. Second draft. http://www.ims.uni-stuttgart.de/projekte/mate/mdag/cr/cr_1.html.

Simard Michel (1998) The BAF: A Corpus of English-French Bitext. *Proceedings of LREC 98*, Granada, España.

Trouilleux F. (2001) *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*, Thèse de Doctorat, Université Blaise Pascal, Clermont-Ferrand.

Tutin A., Trouilleux F., Clouzot C., Gaussier E., Zaenen A., Rayot S., Antoniadis G. (2000) Annotating a large corpus with anaphoric links, In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)*. Lancaster, UK.

Uehara, S. (1996) "Anaphoric pronouns in English and their counterparts in Japanese". *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC96)*. Lancaster, UK. 64-75.