Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns

Andrei Popescu-Belis*, Thomas Meyer*, Jeevanthi Liyanapathirana*, Bruno Cartoni†, Sandrine Zufferey‡

*Idiap Research Institute
Rue Marconi 19, 1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch, thomas.meyer@idiap.ch, juliyanapathirana@gmail.com

†Department of Linguistics, University of Geneva
2, rue de Candolle, 1211 Geneva 4, Switzerland
bruno.cartoni@unige.ch

‡Institut Langage et Communication, Université catholique de Louvain
Place Blaise Pascal, 1 – L3.03.33, 1348 Louvain-la-Neuve, Belgique
sandrine.zufferey@gmail.com

Abstract

This paper describes methods and results for the annotation of two discourse-level phenomena, connectives and pronouns, over a multilingual parallel corpus. Excerpts from Europarl in English and French have been annotated with disambiguation information for connectives and pronouns, for about 3600 tokens. This data is then used in several ways: for cross-linguistic studies, for training automatic disambiguation software, and ultimately for training and testing discourse-aware statistical machine translation systems. The paper presents the annotation procedures and their results in detail, and overviews the first systems trained on the annotated resources and their use for machine translation.

Keywords: discourse connectives, pronouns, machine translation

1. Motivation

Discourse-level phenomena remain a challenge for statistical machine translation (SMT), as current SMT models are mainly at the phrase-level, whereas the correct translation of several types of discourse phenomena requires modeling over multiple sentences or paragraphs. Among these phenomena, the COMTIS project targets SMT improvement on discourse connectives, pronouns, and verb tenses, first on the English/French pair, and then including German and Italian.[1]

The main idea behind COMTIS is that natural language processing (NLP) techniques can be used to solve discourse-level ambiguities with sufficient accuracy to improve SMT output, by making new features available to MT. Therefore, a key element is the availability of discourse-annotated resources to train and test NLP components and combined NLP+MT systems.

This paper presents challenges and solutions for the annotation of discourse connectives and pronouns for the specific purpose of MT in COMTIS, along with the resulting resources and their use. The discussion will bear on the English/French pair of the Europarl Corpus of transcribed EU parliamentary debates (Koehn, 2005), using directional subsets that were originally produced in the source language (Cartoni and Meyer, 2012).[2]

This paper is organized as follows. The two main sections (Section 2 and 3) are dedicated to the two discourse-level phenomena addressed in this paper, i.e. discourse connectives and pronouns. In both cases, existing resources are summarized, followed by the presentation of the annotation method we adopted, and the new resources thus created. Section 4 sums up the results of the annotation and presents how they are used within the COMTIS project.

2. Annotation of Discourse Connectives

Discourse connectives are words or phrases that make explicit the functional or rhetorical relations between propositions. The main challenge they raise for translation is the ambiguity of many frequent connectives, which may signal different relations upon different uses. For instance, in English, *while* may signal temporal relations, but also concessions, contrasts, or comparisons. The translation of an occurrence of *while* into French generally depends on its sense. Finding the exact sense of a connective often requires non-local information, such as the presence in the sentence of contrastive terms, or on local but connective-specific features, both of which are difficult to learn by current SMT systems. This is why an indication of the exact sense of a connective, produced by an NLP component prior to SMT, is potentially helpful for translation. To train

[1] COMTIS stands for Improving the Coherence of Machine Translation Output by Modeling Intersentential Relations, see www.idiap.ch/comtis. COMTIS is a project supported by the Swiss NSF in the Sinergia program.

[2] The directional excerpts from Europarl are made available at www.idiap.ch/dataset/europarl-direct/.

and evaluate such an NLP component, language resources annotated for discourse connectives are necessary.

2.1. Classic Approach: Annotation of the Connective Meaning

The sense-based approach to annotation is exemplified by the English Penn Discourse Treebank (PDTB) (Prasad et al., 2008), or its more recent Czech counterpart (Zikánová et al., 2010). The PDTB annotation manual defined a hierarchy of relation types between clauses or propositions, which the annotators applied to relations signaled explicitly by discourse connectives, but also to implicit ones, using one or more labels per relation. Of course, in this approach, the annotation manual must define applicability conditions as clearly as possible. However, especially when sense labels are very detailed (to map fine-grained linguistic analyses, for instance), the inter-annotator agreement is quite low, and the annotation process becomes very costly. In our experiments, when five annotators annotated the connective *while* with one of four possible labels,[3] the average *kappa* measure was only 0.56 over 30 pilot sentences.
It appears however that fine-grained sense-based labeling of relations and/or connectives is not a necessity for MT, and a more efficient method oriented towards MT was translation spotting, as explained in the following section.

2.2. Annotation with Translation Spotting Approach

In translation spotting, as explained in (Cartoni et al., 2011), human annotators are asked to annotate on sentence pairs the exact translation of each source connective into the target language.[4] The main motivation for adopting the approach is to rely on the judgment of the translator, who, when translating the text, had access to the full text and made choices according to the meaning of the connective to be translated.
Figure 1 provides examples of English sentences containing the connective *since*, their parallel translated sentences (extracted from the Europarl corpus), and the 'transpot', i.e. the manually spotted translation of the connective. In most cases, the translation uses a target language connective signaling one of the possible senses of the source connective, but reformulations and the absence of an explicit connective are also possible. For instance, in Example 4 in Figure 1, the connective *since* has been translated by a paraphrase (P). The second step consists of clustering all similar translations of each connective type, and labeling the clusters with *a posteriori* sense labels. In the examples in Figure 1, *since* in sentences 1 and 2 is translated respectively by the connectives *étant donné que* and *puisque*, which have a causal meaning in French, while in the two other sentences, the transpot indicates a temporal meaning. From this translation spotting and clustering steps, the meaning of *since* can be inferred and annotated accordingly.
The advantages of this method with respect to a predefined annotation scheme including a list or hierarchy of labels, given our purpose, are:

1. a simpler resulting sense scheme, more tractable for automatic labeling;
2. empirical grounding in a given type of data;
3. adaptation to the translation problem (the labels are those that make a difference in translation);
4. significant speedup of the annotation.

But the translation spotting approach also has some drawbacks:

1. specificity to a given language pair;
2. different senses that are rendered by the same connective in translation are not distinguished.

Nevertheless, the approach helps to produce large sets of annotated data, as is explained in the following section.

2.3. Created Resources for Discourse Connectives

Several types of discourse connectives, among which the most ambiguous ones, have been processed, aiming at 200 occurrences or more per type, and results are shown in Table 1. These types were selected because they were described in monolingual studies as having multiple possible senses – e.g. in various dictionaries, the PDTB, or LexConn (Roze et al., 2008). When annotating using translation spotting, the *a posteriori* senses were sometimes different from the principal *a priori* ones listed in the literature, and both lists are represented in Table 1. Some sentences were discarded due to non-connective uses or other problems due to the automatic extraction of the occurrences. A total of 3231 connectives (2514 English and 817 French), of 12 types (8 English and 4 French), have been annotated, and more fine-grained annotations of causal connectives are under work.

3. Annotation of Pronouns

To improve SMT for pronouns – as in the work of Le Nagard and Koehn (2010) or Hardmeier and Federico (2010) for instance – annotated resources are again necessary. The translation of a pronoun is governed, among other factors, by discourse constraints that are mainly due to the nature of its antecedent, i.e. the referent it stands for. Therefore, a local approach to translation is likely to make wrong decisions, for instance when assigning the gender of a target-language pronoun when the source-language pronoun has no gender marking.
There is however a conceptual problem regarding what needs to be annotated, because the translation of a pronoun depends on the translation of its antecedent, and this translation is not known before actually submitting the antecedent to machine translation. For instance, when the target language marks grammatical gender but the source one does not, a source pronoun cannot (in theory) be annotated for certain with the gender it should have in translation, because the exact wording of the translated antecedent (hence, its grammatical gender) is not fixed a priori and is a matter of translation choice. Depending on the language pair, additional features related to the antecedent or to discourse properties can play a role in the translation of a pronoun, for instance humanness, emphasis, or level of politeness.

[3] These were: 'temporal', 'concession', 'comparison' and 'contrast', and had been defined and exemplified in the annotation instructions.

[4] This can also be done automatically, but with insufficient precision (Danlos and Roze, 2011).

	English Sentence	French Sentence	Transpot
1	In this regard the technology feasibility review is necessary, **since** the emission control devices to meet the ambitious NOx limits are still under development.	À cet égard, il est nécessaire de mener une étude de faisabilité, **étant donné que** les dispositifs de contrôle des émissions permettant d'atteindre les limites ambitieuses fixées pour les NOx sont toujours en cours de développement.	étant donné que
2	Will we speak with one voice when we go to events in the future **since** we now have our single currency about to be born?	Parlerons-nous d'une seule voix lorsque nous en arriverons aux événements futurs, **puisqu**'à présent notre monnaie unique est sur le point de voir le jour?	puisque
3	In East Timor an estimated one-third of the population has died **since** the Indonesian invasion of 1975.	Au Timor oriental, environ un tiers de la population est décédée **depuis** l'invasion indonésienne de 1975.	depuis
4	It is two years **since** charges were laid.	**Cela fait** deux ans **que** les plaintes ont été déposées.	P (cela fait X que)

Figure 1: Examples of parallel sentences with the English connective *since* and its translation spotting in French. In the fourth example, the translation is not an explicit connective, but a paraphrase.

	English Sentence	French Sentence	Transpot
1.	**It** cannot create jobs just by government spending.	**Elle** ne peut créer des emplois uniquement par le biais de dépenses gouvernementales.	elle
2.	**It** says that this should be done despite the principle of relative stability.	**Il** précise que cela devrait être fait malgré le principe de stabilité relative.	il
3.	So far as the September communication is concerned, **it** is unprecedented.	S'agissant de la communication de septembre, on ne **lui** connaît aucun précédent.	lui
4.	**It** is fundamentally essential therefore that the authority is not anonymous.	**Il** est par conséquent essentiel que cette autorité ne soit pas anonyme.	il

Figure 2: Examples of parallel sentences with the English pronoun *it* and the corresponding translation spotting in French. See Section 3.2 for a discussion of the actual antecedents of the pronouns.

3.1. Classic Approach: Annotation of the Pronoun Antecedent

Annotation of pronouns for MT could thus amount to annotating their antecedents, as in the previous studies cited above, i.e. it would seem that the only reasonable candidate for annotation is simply the antecedent of the pronoun. Resources annotated for anaphora or coreference are in fact already available – e.g. the MUC-6 and MUC-7 corpora, or the OntoNotes English Coreference data, see for instance (Recasens and Hovy, 2010), or ELDA catalog n. ELRA-W0032 and LDC catalog n. LDC2003T13, LDC2001T02 and LDC2011T01 – mainly in English and French, but not over parallel corpora.

The main problem of the antecedent-oriented approach is that using such monolingual resources for training and testing SMT requires first the identification of the candidate translation of each antecedent, which might be difficult, and then presupposes that when testing, the MT system will perform anaphora resolution. However, state-of-the-art scores for anaphora resolution are still quite low, as indicated by Le Nagard and Koehn (2010) or Hardmeier and Federico (2010). In our own assessment over 45 EN sentences from Europarl, we found that four freely available systems gave on average about 40% correct answers only.

3.2. Proposed Approach: Annotation of a Pronoun's Reference Translation (Transpot)

We have explored an alternative solution which annotates the reference translation of a pronoun without annotating its antecedent. As mentioned above, several correct translations of the pronoun are acceptable, in theory, depending on the antecedent's translation, but in practice, from English to French, the observed range of possible translations is very narrow.[5] We have thus applied the translation spotting method to the annotation of the English pronoun *it* into French (intended for use in MT). While in English *it* refers to non-human entities, its French translation depends on the grammatical gender of the referent, which is absent from English, as well as on the pleonastic or emphatic role and some idiosyncratic constructions. The a priori possible translations, listed also in Table 1, are: *il, elle, le, la, l', lui, cela, celui, celui-ci, celle-là, ce, c', en,* and *y*.

3.3. Created Resources for Pronouns

The resource produced includes nearly 400 instances of *it* annotated using translation spotting on the French transla-

[5]This may be due to the fact that many pronoun features (e.g. number and to a certain extent gender) depend on the referent itself, not from the referring expression.

Lexical items	A priori senses	A posteriori senses	N.S.	F.S.
EN CONNECTIVE			*Total EN: 2,379*	
as		preposition; connective: causal, concession, comparison, temporal	600	599
although	contrast, concession	contrast, concession	197	183
even though	contrast, concession	contrast, concession	212	191
meanwhile	contrast, temporal	contrast, temporal	131	131
since	temporal, causal	temporal, temporal_and_causal, causal_known_relation, causal_new_relation, causal_other	558	558
though	contrast, concession	contrast, concession	200	155
while	contrast, concession, comparison, temporal	contrast, concession, contrast_and_ temporal, temporal_durative, temporal_punctual, temporal_causal	499	294
yet	adverb, contrast, concession	adverb, contrast, concession	509	403
FR CONNECTIVE			*Total FR: 817*	
alors que	contrast, temporal	contrast, temporal, temporal_and_contrast	423	366
bien que	concession	contrast, concession	55	51
dans la mesure où	condition, explanation	condition, explanation	175	150
pourtant	contrast, concession	contrast, concession	312	250
EN PRONOUN				
it	*il, elle, le, la, l', lui, cela, celui, celui-ci, celle-là, ce, c', en, y*	*il, elle, le, la, lui*	393	393

Table 1: List of created resources in English and French. N.S. stands for number of automatically-extracted sentences submitted to annotators, and F.S. for the number of final sentences retained. The *a priori* senses, when indicated, are based on the PDTB or LexConn labels, while the *a posteriori* ones, as explained in the text, were defined by clustering after translation spotting and are specific to this work. Two sense labels clustered with '_and_' reflect the case when the sense distinction is not relevant to translation, i.e. both senses can be conveyed by the same connective in the target language.

tion. These instances include uses of pleonastic or "impersonal" (non-referential) *it*. For each occurrence of *it*, one annotator selected the French translation, finding out a posteriori that a majority of translations of *it* were by: *il, elle, lui, le, la* (or *l'*).

Figure 2 exemplifies the translation spotting of the English pronoun *it* in French for four different cases. In sentence 1, the referent of the pronoun is *Europe* (in a sentence before this one) which is feminine in French and which is why the correct pronoun here is *elle*. The second example is a similar case, but here the referent in a sentence before is *the paragraph* (translated by *le paragraphe* in French) and therefore *it* is translated by the masculine French pronoun *il*. In the third example, the translation of *it* in French is case-dependent (indirect object) and therefore has to be *lui*, which is both masculine and feminine; the referent is here *the communication*, translated by *la communication* in French. In example 4, the use of *it* is pleonastic, i.e. impersonal or non-referential, and the construct *it is essential* translates to *il est essentiel* with the pleonastic pronoun *il* in French. In our observations, a large number of pleonastic constructions in English are translated by similar constructions in French.

Such an annotation method has two biases for pronouns, which have a yet undetermined effect on subsequent MT scores. Firstly, as discussed, the annotation presupposes that all correct translations of the antecedent will have the same grammatical features as the reference translation. This fact was not contradicted by our data: all correct translations produced by a baseline Moses SMT system (Koehn et al., 2007) trained on Europarl had the same gender as the reference translation. Secondly, if the annotation is used as is for evaluation, pronoun translations will be considered as correct only when they are identical to the reference translation, thus displaying a similar behavior to the BLEU measure (Papineni et al., 2002) with one reference.

4. Synthesis of Results and their Use

Table 1 provides a summary of the resources that have been annotated so far through translation spotting. The annotated resources have been used for training and testing automatic labeling modules, the output of which has then been piped into a statistical MT system, with the goal of improving translations.

The annotation of *discourse connectives* has served for initial experiments with automatic labeling of connectives (Meyer, 2011; Meyer et al., 2011), which showed competitive results in comparison to previous studies (Pitler and Nenkova, 2009). In more recent work, the classifiers have become more accurate due to more semantically oriented features and the combination of the connective disambiguation modules with SMT systems demonstrated a small improvement in global MT quality (Meyer and Popescu-Belis, 2012), measured by BLEU, despite the overall scarcity of

connectives.

Classifiers built for pronoun correction were trained and tested on this annotation using 5-fold cross validation. The goal is to predict, given a candidate translation of a pronoun and a set of features extracted from its sentence and the preceding one, whether the candidate translation is correct, or whether it should be chnaged, and how (for instance, an *il* changed into *elle* or vice-versa). Among the features that were considered was information about the gender of the preceding nouns in the same and previous sentence (e.g., majority, most recent, etc.) along with the candidate translation of the pronoun (determined from a GIZA++ alignment), and positional and grammatical features of the pronoun, candidate, and neighboring words. The accuracy of the best C4.5 decision trees was around 60%. Overall, this improved pronoun choice (with respect to a baseline Moses SMT) from incorrect to correct in about 27% of the cases, but also degraded it in 16% of the cases. The BLEU score showed a small but significant improvement.

5. Conclusion and Perspectives

This paper presented an annotation method of two discourse-level phenomena, connectives and pronouns, intended for producing resources for discourse-aware machine translation. The method involves translation spotting and clustering over a parallel corpus. Results of human annotation and machine labeling showed that this method was more tractable than explicit sense annotation, and provided results that were more relevant to our objective. The resulting resources also led to a small but significant improvement of the MT output. Future developments will include the addition of a wider range of connectives and pronouns and the annotation of verb tenses. A more detailed evaluation of the MT improvement brought by each of the resources is also under way.

Acknowledgments

We are grateful for the funding of this work to the Swiss National Science Foundation (SNSF), under its Sinergia program, grant n. CRSI22_127510. The resources described in this article will be made available through the project's website (`www.idiap.ch/comtis`) in the near future.

6. References

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. *4th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 78–86, Portland, OR.

Bruno Cartoni and Thomas Meyer. 2012. Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. *LREC 2012*, Istanbul.

Laurence Danlos and Charlotte Roze. 2011. Traduction (automatique) des connecteurs de discours. *TALN 2011*, Montpellier.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. *IWSLT 2010*, Paris.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. *ACL 2007 Demonstration Session*, pages 177–180, Prague.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit X*, pages 79–86, Phuket.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. *Workshop on SMT and Metrics*, pages 258–267, Uppsala.

Thomas Meyer. 2011. Disambiguating temporal-contrastive discourse connectives for machine translation. *ACL-HLT 2011 Student Session*, pages 46–51, Portland, OR.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. *SIGdial 2011*, pages 194–203, Portland, OR.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using Sense-labeled Discourse Connectives for Statistical Machine Translation. *ESIRMT-HyTra Workshop at EACL 2012*, Avignon, France.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *ACL 2002*, pages 311–318, Philadelphia, PA.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. *ACL-IJCNLP 2009 Short Papers*, pages 13–16, Singapore.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *LREC 2008*, pages 2961–2968, Marrakesh.

Marta Recasens and Eduard Hovy. 2010. Coreference Resolution across Corpora: Languages, Coding Schemes, and Preprocessing Information. *ACL 2010*, pages 1423–1432, Uppsala.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2010. LEXCONN: a French Lexicon of Discourse Connectives. *Workshop on Multidisciplinary Approaches to Discourse (MAD)*, pages 114-125, Moissac, France.

Zikánová, Sárka and Mladová, Lucie and Mírovský, Jiří and Jínová, Pavlina. 2010. Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank. *LREC 2010*, pages 2002–2006, Valletta.