Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation

Thomas Meyer and **Andrei Popescu-Belis**
Idiap Research Institute
Rue Marconi 19, 1920 Martigny, Switzerland
Thomas.Meyer@idiap.ch, Andrei.Popescu-Belis@idiap.ch

Sandrine Zufferey and **Bruno Cartoni**
Department of Linguistics, University of Geneva
Rue de Candolle 2, 1211 Geneva 4, Switzerland
Sandrine.Zufferey@unige.ch, Bruno.Cartoni@unige.ch

Abstract

Many discourse connectives can signal several types of relations between sentences. Their automatic disambiguation, i.e. the labeling of the correct sense of each occurrence, is important for discourse parsing, but could also be helpful to machine translation. We describe new approaches for improving the accuracy of manual annotation of three discourse connectives (two English, one French) by using parallel corpora. An appropriate set of labels for each connective can be found using information from their translations. Our results for automatic disambiguation are state-of-the-art, at up to 85% accuracy using surface features. Using feature analysis, contextual features are shown to be useful across languages and connectives.

1 Introduction

Discourse connectives are generally considered as indicators of discourse structure, relating two sentences of a written or spoken text, and making explicit the rhetorical or coherence relation between them. Leaving aside the cases when connectives are only implicit, the presence of a connective does not unambiguously signal a specific discourse relation. In fact, many connectives can indicate several types of relations between sentences, i.e. they have several possible "senses" in context.

This paper studies the manual and automated disambiguation of three ambiguous connectives in two languages: *alors que* in French, *since* and *while* in English. We will show how the multilingual perspective helps to improve the accuracy of annotation, and how it helps to find appropriate labels for automated processing and MT. Results from automatic annotation experiments, which are close to the state of the art, as well as feature analysis, help to assess the usefulness of the proposed labels.

The paper is organized as follows. Section 2 explains the motivation of our experiments, and offers a wider perspective on our research goals, illustrating them with examples of translation problems which arise from ambiguous discourse connectives. Current resources and methods for discourse annotation are discussed in Section 3. Section 4 analyzes our experiments in manual annotation and in particular the influence of the set of labels on the reliability of annotation. The automatic disambiguation experiments, the features used, the results and the analysis of features are described in Section 5. Section 6 concludes the paper and outlines future work.

2 Explicit Connectives and their Translation

2.1 Three Multi-functional Connectives

Discourse connectives form a functional category of lexical items that are used to mark coherence relations such as *Cause* or *Contrast* between units of discourse. Along with other function words, many connectives appear among the most frequent words, as shown for instance by counts (Cartoni et al., 2011) over the Europarl corpus (Koehn, 2005). The Penn Discourse Treebank (Prasad et al., 2008) (see Section 3.1 below) includes around 100 connective types, but the exact number varies across studies,

depending on the discourse theory used to classify them. Among these types, Pitler et al.(2008) have shown that most of them are unambiguous and easy to identify, but others, especially temporal ones, often signal multiple senses depending on their context.

Following the terminology of Petukhova and Bunt (2009, Section 2), we are interested here in "sequential" multi-functionality, i.e. the fact that the same connective can signal different relations in different contexts. We do not deal with "simultaneous" multi-functionality, i.e. the possibility for a single occurrence to signal several relations, which has been less frequently studied for connectives (see Petukhova and Bunt (2009) for the discourse usage of *and*).

We identified the two English connectives *while* and *since*, along with the French connective *alors que*, as being particularly problematic because they are highly multi-functional, i.e. they can signal multiple senses. For *alors que*, a French database of connectives (LexConn (Roze et al., 2010), see Section 3 below) contains examples of sentences where *alors que* expresses either a *Background* or a *Contrast* relation. For the English connective *since*, Miltsakaki et al. (2005) identified three possible meanings: *Temporal*, *Causal*, and simultaneously *Temporal/Causal*. For *while*, even more senses are observed: *Comparison*, *Contrast*, *Concession*, and *Opposition*. In fact, in the Penn Discourse Treebank, the connective *while* is annotated with more than twenty different senses.

2.2 Wider Research Objectives

Our long-term goal is to identify automatically the senses of connectives for an application to machine translation (MT). Going beyond the labels provided by discourse theories, the goal is thus to find the most appropriate labels in a new multilingual, empirical approach that makes use of parallel corpora to annotate and then learn the various senses of connectives. The disambiguation of such connectives in a source text is crucial for its translation, because each sense may be translated by a different connective and/or syntactical construct in the target language.

More specifically, we hypothesize that correctly labeled connectives are easier to learn and to translate by statistical MT systems than unlabeled ones. To support this hypothesis, we set up an experiment (Meyer, 2011) in which we constrained the translation of the three senses of the discourse connective *while* that were previously annotated as *Temporal*, *Contrast* and *Concession*. The system was forced to use predefined French translations known to be correct, by directly modifying the phrase table of the trained MT system. This modification noticeably helped to improve translation quality and rose the BLEU score by 0.8 for a preliminary test set of 20 sentences.

2.3 Illustration of Mistranslations

Among the connectives that we plan to process in order to improve MT, the three connectives we focus on in this paper are frequent, ambiguous and therefore difficult to translate correctly by MT systems, as illustrated in the following examples.

A first reason why machine translation of connectives can be difficult is that there may be no direct lexical correspondence for the explicit source language connective in the target language, as shown in the reference translation of the first example in Table 1, taken from the Europarl corpus (Koehn, 2005).

EN	*It is also important that we should not leave these indicators floating in the air* ***while*** *congratulating ourselves on the fact that we have produced them.*
FR	*Il est également important de ne pas laisser ces indicateurs flotter,* ***en*** *nous félicit****ant*** *de les avoir instaurés.*
EN	*Finally, and in conclusion, Mr President, with the expiry of the ECSC Treaty, the regulations will have to be reviewed* ***since*** *[causal] I think that the aid system will have to continue beyond 2002 . . .*
FR	**Enfin, et en conclusion, Monsieur le président, à l'expiration du traité ceca, la réglementation devra être revu* ***depuis que*** *[temporal] je pense que le système d'aides devront continuer au-delà de 2002 . . .*
FR	*Oui, bien entendu, sauf que le développement ne se négocie pas,* ***alors que*** *[contrast] le commerce, lui, se négocie.*
EN	**Yes, of course, but development cannot be negotiated,* ***so*** *[causal] that trade can.*
EN	*Between 1998 and 1999, loyalists assaulted and shot 123 people,* ***while*** *[contrast] republicans assaulted and shot 93 people.*
FR	**Entre 1998 et 1999, les loyalistes ont attaqué et abattu 123 personnes, ϕ 93 pour les républicains.*

Table 1: Translation examples from Europarl. Discourse connectives, their translations, and their senses are indicated in bold. The first example is a reference translation from EN into FR, while the others are wrong translations generated by MT (EN/FR and respectively FR/EN), hence marked with an asterisk.

When an ambiguous connective is explicitly translated by another connective, the incorrect rendering of its sense can lead to erroneous translations, as in the second and third examples in Table 1, which are translated by the Moses SMT decoder (Koehn et al., 2007) trained on the Europarl corpus. The reference translation for the second example uses the French connective *car* with a correct causal sense, instead of the wrong *depuis que* generated by SMT, which expresses a temporal relation. In the third example, the French connective *alors que*, in its contrastive usage, is wrongly translated into the English connective *so*, which has a causal meaning (the reference translation uses *whereas* to express contrast). It may even occur that the system fails to translate a connective at all, as in the fourth example where the discourse information provided by *while*, namely a *Contrast* relation, is lost in the French translation, which is hardly coherent any longer.

3 Related Work

3.1 Annotated Resources

One of the very few available discourse annotated corpora is the Penn Discourse Treebank (PDTB) in English (Prasad et al., 2008). For this resource, one hundred types of explicit discourse connectives were manually annotated, as well as implicit relations not signaled by a connective. The sense hierarchy used for annotation consists of three levels, from four top-level senses (*Temporal*, *Contingency*, *Comparison*, and *Expansion*), to 16 subsenses on the second level, and 23 further ones on the third level. The annotators were allowed to assign more than one sense to each occurrence, so 129 simple or complex labels are observed, over more than 18,000 explicit connectives. For French, the ANNODIS project (Péry-Woodley et al., 2009) will provide annotation of discourse on an original corpus. Resources for Czech are also becoming available (Zikánová et al., 2010).

For German, a lexicon of discourse markers named DiMLex exists since the 1990s (Stede and Umbach, 1998). An equivalent, more recent database for French is the LexConn lexicon of connectives (Roze et al., 2010) containing a list of 328 explicit connectives. For each of them, LexConn indicates and exemplifies the possible senses, chosen from a list of 30 labels inspired from Rhetorical Structure Theory (Mann and Thompson, 1988).

3.2 Automatic Disambiguation of Connectives

The release of the PDTB had quite an impact on automatic disambiguation experiments. The state-of-the-art for recognizing all types of explicit connectives in English is therefore already high, at 97% accuracy for disambiguating discourse vs. non-discourse uses (Lin et al., 2010) and 94% for disambiguating the four main senses from the PDTB hierarchy (Pitler and Nenkova, 2009). Lin et al. (2010) recently built the first end-to-end PDTB discourse parser, which is able to parse unrestricted text with an F1 score of 38.18% for senses on the second level of the PDTB hierarchy. Other important contributions to automatic discourse connective classification and feature analysis has been provided by Wellner et al. (2006) and Elwell and Baldrige (2008).

Fewer studies focus on the detailed analysis of specific discourse connectives. In Section 5.3, we will compare our results to Miltsakaki et al. (2005) who report classification results for the connectives *since*, *while* and *when*. In their study, as in the present one, the goal is to disambiguate senses from the second level of the PDTB hierarchy, a level which, as we will show, is appropriate for the translation of these connectives as well.

4 Connective Annotation in Parallel Corpora

The resources mentioned above are either monolingual only (PDTB, LexConn) and/or not yet publicly available (ANNODIS, DiMLex). Moreover, our overall goal is related to multilingualism and translation, as explained in Section 2.2 above. Therefore, we performed manual annotation of connectives in a multilingual, aligned resource: the Europarl corpus (Koehn, 2005). We extracted from Europarl two subcorpora for each translation direction, EN/FR and FR/EN, to take into account the varying distribution of connectives in translated vs. original language, as explained in Cartoni et al. (2011).

As the full PDTB hierarchy seemed too fine-grained given current capabilities for automatic labeling and the needs for translating connectives, we defined a simplified set of labels for the senses of connectives, by considering their usefulness and

granularity with respect to translation, focusing on those that may lead to different connectives or syntactical constructs in the target language.

4.1 Method

There are two major ways to annotate explicit discourse connectives. The first approach is to label each occurrence of a connective with a label for its sense, similar to the PDTB or LexConn hierarchies of senses. However, as shown among others by Zikanova et al. (2010), this is a difficult and time-consuming task even when the annotators are trained over a long period of time. This is confirmed by the rather low kappa scores resulting from the manual sense annotations as can be seen for each connective in detail below.

The second approach to annotation, which is the one put forward in this paper, is based on *translation spotting*. In a first step, human annotators work on bilingual sentence pairs, and annotate the translation of each connective in the target language. The translations are either a target language connective (signaling in principle the same sense(s) as the source one), or a reformulation, or a construct with no connective at all. In a second step of the annotation, all translations of a connective are manually clustered by the experimenters to derive sense labels, by grouping together similar translations.

As demonstrated in the following subsections, for the three connectives under study, the second approach to connective annotation not only facilitates the annotation task, but also helps to derive the appropriate level of granularity for the sense labels.

4.2 Annotation of *alors que*

This first manual annotation involved two experienced annotators who annotated *alors que* in 423 original French sentences. The two main senses identified for *alors que* are *Background* (labeled B) *Contrast* (labeled C), as in the LexConn database. Annotators were also allowed to use the J label if they did not know which label to assign, and a D label for discarded sentences – due to a non-connective use of the two words which could not be filtered out automatically (e.g. *Alors, que fera-t-on?*). The annotators found 20 sentences labeled with D, which were removed from the data. 15 sentences were labeled with J by one annotator (but none by both), and it was decided to assign to them the label (either B or C) provided by the other annotator.

The inter-annotator agreement on the B vs. C labels was quite low, showing the difficulty of the task: *kappa* reached 0.43, quite below the 0.7 mark often considered as indicating reliability. The following example from Europarl illustrates the difficulty of choosing between B and C. In particular, the reference translation into English also uses an ambiguous connective, namely *while*.

FR *La monnaie unique va entrer en vigueur au milieu de la tourmente financière,* ***alors que*** *de nombreux compléments, logiques, mais que les États ne semblaient pas avoir prévus, n'ont pas encore été apportés.*

EN *The single currency is going to come into force in the midst of financial turmoil,* ***while*** *a great many additional factors which were only to be expected, but which the states do not seem to have anticipated, have not been taken into consideration.*

Two methods were applied to deal with diverging manual annotations. To prepare the datasets for the automated disambiguation experiments, one solution (named A1, see Table 2) is to use the double-sense label B/C for sentences labeled differently by annotators (B vs. C). This label reflects the difficulty of manual annotation and preserves the ambiguity which might be genuinely present in each occurrence. The relevance of the B/C label is also supported by results from automatic labeling in Section 5.3 below.

For comparison purposes, a second dataset named A2 was derived from translation spotting on the same French sentences aligned to English ones, as explained in Section 4.1. *Alors que* appeared to be mainly translated by the following English equivalents and constructs: *although*, *whereas*, *while*, *whilst*, *when*, *at a time when*. Through this operation, inter-annotator disagreement can sometimes be solved: when the translation is a clearly contrastive English connective (*whereas* or *although*), then the C label was assigned instead of B/C. Conversely, when the English translation was still ambiguous (*while*, *whilst*, or *when*), the experimenters made a decision in favor of either B or C by re-examining source and target sentences.

4.3 Annotation of *since*

For *since*, 30 sentences were annotated by four experimenters in a preliminary round, with a kappa

ID	Connective	Sent.	Labels (nb. of occ.)
A1	alors que	403	B (92), C (191), B/C (120)
A2	alors que	403	B (126), C (277)
B1	since	727	T (375), C (341), T/C (11)
B2	since	727	T (375), C (352)
C1	while	299	T/C (92), CONC (134), C (43) T/CAUSAL (19), T/DUR (7) T/PUNCT (4)
C2	while	299	T (30), C (135), CONC (134)

Table 2: The six datasets resulting from the manual annotation of the three connectives, with total number of sentences, possible labels and their number of occurrences. The explanations of the labels are given in Sections 4.2 through 4.4.

score of 0.77, indicating good agreement. Then, each half of the entire dataset (727 sentences) was annotated by another person with three possible sense labels: T for *Temporal*, C for *Causal* and T/C for a simultaneously *Temporal/Causal* meaning. Two datasets were again derived from this manual annotation. To study the effects of a supplementary label, we kept the label T/C for dataset B1, but condensed it under label C in dataset B2, as shown in Table 2.

4.4 Annotation of *while*

The English connective *while* is highly ambiguous. In the PDTB, occurrences of *while* are annotated with no less than 21 possible senses, ranging from *Conjunction* to *Contrast*, *Concession*, or *Synchrony*.

We performed a pilot annotation of 30 sentences containing *while* with five different experimenters, resulting in a quite low inter-annotator agreement, $\kappa = 0.56$. We therefore decided to perform a translation spotting task only, with two experienced annotators fluent in English and French. The observed translations into French confirm the ambiguity of *while*, as they include several connectives and constructs, quite evenly distributed in terms of frequency: *alors que*, gerundive reformulations, other reformulations, *si*, *tandis que*, *même si*, *bien que*, etc.

The translations were manually clustered to derive senses for *while*, in an empirical manner. For example, *alors que* signals *Temporal/Contrast*, which is also true for *tandis que*. Similarly, *même si* and *bien que* are clustered under the label *Concession*, and so forth. The translation spotting shows that at least *Contrast*, *Concession*, and several temporal senses are necessary to account for a correct translation. These distinctions are comparable to the semantic granularity of the second PDTB hierarchy level.

To generate training sets for automated classification out of a total of 500 sentences, we discarded 201 sentences labeled by annotators with G (gerundive constructions), P (reformulations) or Z (no translation at all) – these cases could be reconsidered in further work, as they represent valid translation problems. For the remaining 299 sentences, we created the following six labels by clustering the spotted translations: T/C (*Temporal/Contrast*), T/PUNCT (*Temporal/Punctual*), T/DUR (*Temporal/Duration*), T/CAUSAL (*Temporal/Causal*), CONC (*Concession*) and C (*Contrast*). These were used to tag the remaining 299 sentences, forming dataset C1. A second dataset (C2) with fewer senses was obtained from C1 by merging T/C to C (*Contrast* only) and all T/x to T (*Temporal* only).

5 Disambiguation Experiments

The features for connective classification, the results obtained and a detailed feature analysis are discussed in this section. We show that an automated disambiguation system can be used to determine the most appropriate set of labels, and thus to corroborate the selection we made using translation spotting.

5.1 Features

For feature extraction, all the datasets described in Section 4 were processed as follows. The English texts were parsed and POS-tagged by Charniak and Johnson's (2005) reranking parser. The French texts were POS-tagged with the MElt tagger (Denis and Sagot, 2009) and parsed with MaltParser (Nivre, 2003). As the English parser provides constituency trees, and the parser for French generates dependency trees, the features are slightly different in the two languages. The other features below were extracted using elementary pre-processing of the sentences.

For English sentences, we used the following features: the sentence-initial character of the connec-

tive (yes/no); the POS tag of the first verb in the sentence; the type of first auxiliary verb in the sentence (if any); the word preceding the connective; the word following the connective; the POS tag of the first verb following the connective; the type of the first auxiliary verb after the connective (if any).

For French sentences, the features were the following: the sentence-initial character of the connective (yes/no); the dependency tag of the connective; the first verb in the sentence; its dependency tag; the word preceding the connective; its POS tag; its dependency tag; the word following the connective; its POS tag; its dependency tag; the first verb after the connective; its dependency tag.

The cased connective word forms from the corpus were not lower-cased, thus keeping the implicit indication of the sentence-initial character of the occurrence, i.e. whether it starts a sentence or not. The output of the POS taggers was used for neighboring words, but not for the connectives, which almost always received the same tag. Charniak's parser for English provides POS tags which differentiate the verb tenses, such as VBD (past), VBG (gerund), and so on. These were considered for the verb directly preceding and the one directly following the connective. Tense was believed to be potentially relevant because *since* and *while* can have temporal meanings.

The occurrence of auxiliary verbs (*be*, *have*, *do*, or *need*) may give additional indications about temporal relations in the sentence. We therefore used the types of auxiliary verbs as features, including the elementary conjugations, represented for *to be* as: *be_present*, *be_past*, *be_part*, *be_inf*, *be_gerund* – and similarly for the other auxiliary verbs, as in (Miltsakaki et al., 2005).

As shown by Lin et al. (2010), duVerle and Prendinger (2009) or Wellner et al. (2006), the context of a connective is very important. We therefore extracted the words preceding and following each connective, the verbs and the first and the last word of the sentences. These may include numbers, sometimes indicating a numerical comparison, time expressions, or antonyms, which could indicate contrastive relations, such as *rise* vs. *fall* (e.g. *It is interesting to see the fundamental stock pickers scream "foul" on program trading when the markets* ***decline****, while hailing the great values still abounding as the markets* ***rise***.).

For French, we likewise extracted the words immediately preceding and following each connective, supplemented by their POS tags. In contrast to constituents, dependency structures contain information about the grammatical function of each word (heads) and link the dependents belonging to the same head. However, as the dependency parser provides no differentiated verb tags, we extracted the verb word forms themselves and added their dependency tags. The same applies to the connective itself, and preceding and following words and their dependency tags.

The dependency tag of the non-connectives varies between *subj* (subject), *det* (determiner), *mod* (modifier) and *obj* (object). The first verb in the sentence often belongs to the *root* dependency while the verb following the connective most often belongs to the *obj* dependency. For *alors que*, the most frequent dependency tags were *mod_mod* and *mod_obj*, indicating the connective's main function as a modifier of its argument.

5.2 Experimental Setting

Our classification experiments made use of the WEKA machine learning toolkit (Hall et al., 2009) to run and compare several classification algorithms: Random Forest (sets of decision trees), Naive Bayes, and Support Vector Machine. The results are reported with 10-fold cross validation on the entire data for each connective, using all features.

Table 3 lists for each method – including the majority classifier as a baseline – the percentage of correctly classified instances (or accuracy, noted *Acc.*), and the *kappa* values. Significance above the baseline is computed using paired t-tests at 95% confidence. When a score is significantly above the baseline, it is shown in *italics* in Table 3. The best scores for each dataset, across classifiers, are indicated in **boldface**. When these scores were not significantly above the baseline, at least they were never significantly below either.

5.3 Results and Discussion

Overall, the SVM classifier performed best, which may be due to the large number of textual features (3 for EN data and 5 for FR data), as SVMs are known to handle them well (Joachims, 1998; du-

ID	Connective	#	Labels	Baseline	R. Forest		N. Bayes		SVM	
				Acc.	*Acc.*	κ	*Acc.*	κ	*Acc.*	κ
A1	*alors que*	403	B, C, B/C	46.9	*53.1*	*0.2*	***55.7***	***0.3***	*54.2*	***0.3***
A2	*alors que*		B, C	68.7	**69.2**	0.1	68.3	***0.2***	64.7	0.1
B1	*since*	727	T, C, T/C	51.6	*79.8*	*0.6*	*82.3*	***0.7***	***85.4***	***0.7***
B2	*since*		T, C	51.6	*80.7*	*0.6*	*84.0*	***0.7***	***85.7***	***0.7***
C1	*while*	299	T/C, T/PUNCT, T/DUR, T/CAUSAL, CONC, C	44.8	43.2	*0.1*	49.9	***0.2***	***52.2***	***0.2***
C2	*while*		T, C, CONC	43.5	*60.5*	***0.3***	*59.9*	***0.3***	***60.9***	***0.3***

Table 3: Disambiguation scores for three connectives (number of occurrences in the training sets), with two sets of labels each, for various classification algorithms. Accuracy (*Acc.*) is in percentage (%), and *kappa* is zero for the baseline method (majority class). The best scores for each data set are in **boldface**, and scores significantly above the baseline (95% t-test) are in *italics*.

Verle and Prendinger, 2009). The maximum accuracy for *alors que* is 55.7%, for *since* it is 85.7%, and for *while* it is 60.9%. While close to other reported values, there is still potential for improvement in the future.

The analysis of results for each data sets leads to observations that are specific to each connective. The high improvement of over the baseline for A1, as opposed to no improvement for A2, confirms the usefulness of the double-sense B/C label for *alors que*, showing that in this case the three-way classification is probably better adapted to the linguistic properties of *alors que* than a two-way classification. Indeed, *alors que*, just as its frequently spotted translation *while*, is linguistically ambiguous in some contexts (see for instance the example in Section 4.2), in which the temporal and the contrastive meaning are likely to co-exist. In the case of A2, where the labels were forced to B or C only, automatic classifiers do not significantly outperform the baseline. While more elaborate features might help, these low scores can be related to the difficulties of human annotators (Section 4.2), and make a strong case against using a two-label schema for *alors que*.

The features used so far lead to high scores for *since* in datasets B1 and B2. The results are comparable to those from Miltsakaki et al. (2005), who used similar features and labels, though with a Maximum Entropy classifier. Moreover, they provide results for individual connectives, and not, as most of the related work for the PDTB, on the whole set of ca. 100 discourse connective types. However, Miltsakaki et al. (2005) used their own datasets for each connective, which are different from the PDTB, because the PDTB was not available at that time. Our SVM classifier outperforms considerably the Maximum Entropy classifier on the three-way classification task (with T, C, T/C), with an accuracy of 85.4% vs. 75.5%, obtained however on different datasets. For the two-way classification (T, C), again on different datasets, our accuracy of 85.7% is slightly lower than the 89.5% given in Miltsakaki et al. (2005).[1]

For *while*, when comparing C1 to C2, it appears that reducing the number of labels from six to three increases accuracy by 8-10%. This is probably due to the small number of training instances for the labels T/PUNCT and T/DUR in C1 for example. However, even for the larger set of labels, the scores are significantly above baseline (52.2% vs. 44.8%), which indicates that such a classifier might still be useful as input to an MT system, possibly improved thanks to a larger training set. The performance obtained by Miltsakaki et al. (2005) on *while* is markedly better than ours, with an accuracy of 71.8% compared to ours of 60.9% with three labels.

5.4 Feature Analysis

The relevance of features can be measured using WEKA by computing the information gain (IG) brought by each feature to the classification task,

[1] In another experiment (Meyer, 2011), we also applied our classifiers to the PDTB data, with less features however. The results were in the same range as those from Miltsakaki et al. (2005), i.e. 75.3% accuracy for *since* and 59.6% for *while*.

R	Feature	IG	
		A1	A2
1	preceding word	1.12	0.64
2	following verb	0.81	0.51
3	first verb	0.74	0.42
4	following word	0.68	0.23
5	preceding word's POS tag	0.15	0.05
5	first verb's dep. tag	0.14	0.06
5	following word's POS tag	0.19	0.03
8	preceding word's dep. tag	0.10	0.03
8	connective's dep. tag	0.09	0.04
10	following word's dep. tag	0.13	0.013
10	following verb's dep. tag	0.04	0.03
12	sentence initial	0.05	0.001

Table 4: Information gain (IG) of features for French connective *alors que*, ordered by decreasing average ranking (R) in experiments A1 and A2. Features 1–4 are considerably more relevant than the following ones.

R	Feature	IG	
		B1	B2
1	preceding word	0.83	0.75
2	following word	0.56	0.52
3	following verb's POS tag	0.24	0.21
4	type of following aux. verb	0.13	0.12
5	type of first aux. verb	0.11	0.11
6	first verb's POS tag	0.02	0.01
7	sentence initial	0.00	0.00

Table 5: Information gain (IG) of features for EN connective *since*, ordered by decreasing average ranking (R) in experiments B1 and B2.

i.e. the reduction in entropy with respect to desired classes (Hall et al., 2009) – the higher the IG, the more relevant the feature. Features can be ranked by decreasing IG, as shown in Tables 4, 5 and 6, in which ranks were averaged over the first and the second data set in each series.

The tables show that across all three connectives and the two languages, the contextual features are always in the first positions, thus confirming the importance of the context of a connective. Following these are verbal features, which are, for these connectives, of importance because the temporal meanings are additionally established by verbal tenses. POS and dependency features seem the least helpful for disambiguation.

R	Feature	IG	
		C1	C2
1	preceding word	1.02	0.65
2	following word	0.83	0.55
3	type of first aux. verb	0.12	0.07
4	following verb's POS tag	0.16	0.04
5	first verb's POS tag	0.07	0.09
5	type of following aux. verb	0.12	0.05
7	sentence initial	0.08	0.07

Table 6: Information gain (IG) of features for EN connective *while*, ordered by decreasing average ranking (R) in experiments C1 and C2. The first two features are considerably more relevant than the remaining ones.

6 Conclusion and Future Work

We have described a translation-oriented approach to the manual and automatic annotation of discourse connectives, with the goal of identifying their senses automatically, prior to machine translation. The manual annotation of the senses of connectives has been enhanced through parallel corpora and translation spotting. This has lead to tag sets that improved both inter-annotator agreement and automatic labeling, which reached state-of-the-art scores. The analysis of relevant features has shown the utility of contextual information.

To improve over these initial results, we will use more semantic information, such as relations found in WordNet between words in the neighborhood of connectives – e.g. word similarity measures and semantic relations such as antonymy. To generate more training instances of the labels found, manual annotation will continue in order to see whether the senses found through translation spotting can improve automatic disambiguation of many more connectives. The annotation of a large parallel corpus will then help to train disambiguation tools along with statistical MT systems that use their output.

Acknowledgments

We are grateful for the funding of this work by the Swiss National Science Foundation (SNSF) under the COMTIS Sinergia Project, n. CRSI22_127510 (see `www.idiap.ch/comtis/`).

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. In *Proceedings of 4th Workshop on Building and Using Comparable Corpora*, Portland, OR.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL 2005 (43rd Annual Meeting of the ACL)*, pages 173–180, Ann Arbor, MI.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009 (23rd Pacific Asia Conference on Language, Information and Computation)*, pages 110–119, Hong Kong, China.

David duVerle and Helmut Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP)*, pages 665–673, Singapore.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proceedings of ICSC 2008 (2nd IEEE International Conference on Semantic Computing)*, pages 198–205, Santa Clara, CA.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11:10–18.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML 1998 (10th European Conference on Machine Learning)*, pages 137–142, Chemnitz, Germany.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007 (45th Annual Meeting of the ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

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

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. Technical Report TRB8/10, School of Computing, National University of Singapore, Singapore.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: towards a functional theory of text organization. *Text*, 8(3):243–281.

Thomas Meyer. 2011. Disambiguating temporal-contrastive discourse connectives for machine translation. In *Proceedings of ACL-HLT 2011 (49th Annual Meeting of the ACL: Human Language Technologies), Student Session*, Portland, OR.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the TLT 2005 (4th Workshop on Treebanks and Linguistic Theories)*, Barcelona, Spain.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of IWPT 2008 (8th International Workshop on Parsing Technologies)*, pages 149–160, Tokyo, Japan.

Marie-Paule Péry-Woodley, Nicholas Asher, Patrice Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, Philippe Muller, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, Laure Vieu, and Antoine Widlöcher. 2009. Annodis: une approche outillée de l'annotation de structures discursives. In *Proceedings of TALN 2009 (16ème Conférence sur le Traitement Automatique des Langues Naturelles)*, Paris, France.

Volha Petukhova and Harry Bunt. 2009. Towards a multidimensional semantics of discourse markers in spoken dialogue. In *Proceedings of IWCS-8 (8th International Conference on Computational Semantics)*, pages 157–168, Tilburg, The Netherlands.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP), Short Papers*, pages 13–16, Singapore.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of Coling 2008 (22nd International Conference on Computational Linguistics), Companion Volume: Posters*, pages 87–90, Manchester, UK.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008 (6th International Conference on Language Resources and Evaluation)*, pages 2961–2968, Marrakech, Morocco.

Charlotte Roze, Laurence Danlos, and Phillippe Muller. 2010. LEXCONN: a French lexicon of discourse connectives. In *Proceedings of MAD 2010 (Multidis-*

ciplinary Approaches to Discourse), pages 114–125, Moissac, France.

Manfred Stede and Carla Umbach. 1998. DiMLex: a lexicon of discourse markers for text generation and understanding. In *Proceedings of ACL 1998 (36th Annual Meeting of the ACL)*, pages 1238–1242, Montreal, Canada.

Ben Wellner, James Pustejovsky, Catherine Havasi, Roser Sauri, and Anna Rumshisky. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of 7th SIGDIAL Workshop on Discourse and Dialogue*, pages 117–125, Sydney, Australia.

Sárka Zikánová, Lucie Mladová, Jiří Mírovský, and Pavlina Jínová. 2010. Typical cases of annotators' disagreement in discourse annotations in Prague Dependency Treebank. In *Proceedings of LREC 2010 (7th International Conference on Language Resources and Evaluation)*, pages 2002–2006, Valletta, Malta.