Using Simple Word-Alignment Measures to Study Discourse Particles

Jennifer Spenader
Center for Language and Cognition
University of Groningen
j.k.spenader@let.rug.nl

This work demonstrates that simple word-alignment measures from statistical machine translation can help in studying the contribution of discourse particles even though these lexical items often do not have recognizable translation equivalents. The aligned Europarl parallel corpora in Swedish and English was mined to extract examples of the use of the Swedish discourse particle ju. *χ^2 tests were used to determine which unigrams and bigrams were significantly more frequent in the English translations of aligned areas where the particle appeared than in the rest of the corpus. Further, ϕ^2 tests were done on relevant significant co-cccuring unigrams and bigrams to rank them in order of the strength of their relationship with the particle. These tests therefore offer a method by which intuitions obtained through introspective study of a small number of examples can be automatically corroborated with a large number of examples.*

1 Introduction

Traditionally, one of the defining characteristics of discourse particles is their inability to contribute to the truth-conditional meaning of the utterances in which they occur. In consequence, it is challenging to give a precise account of these lexical items. Particles are also often considered to have several functions, though it is often difficult to pinpoint what these functions are. Examining a very large number of examples might reveal patterns not recognizable in a smaller study.

However, the majority of work on discourse particles has involved introspective study, necessarily limiting the number of examples that can be studied, and thus it is not clear how universal the results are. This method is often subjective, and fails to result in the identification of reliable and tangible features that can be used to disambiguate particles with multiple functions, a necessary prerequisite for incorporating information into NLP applications.

The multiple functions of some particles can potentially be illuminated by examining how they are rendered in translation. An ambiguous particle in one language may consistently co-occur with one or more transparent expressions in another language and this felicitous mismatch can be exploited. Aijmer (1996) used a small parallel corpus to study the translation of Swedish discourse particles in English. Using a larger corpus would insure that the results generalize to other data.

By using very large parallel corpora we can verify work such as Aijmer's, as well as perhaps discovering new patterns that can give more insight into the function of some particles. In this work two simple word alignment measures used in statistical machine

translation to determine translation equivalents, the χ^2 and ϕ^2 tests, are used to study the Swedish particle, *ju*. Discourse particles differ greatly from the lexical items for which these statistical measures have earlier been shown to be successful in that it is not clear to what extent co-occurring lexical items can truly be called 'translations' of the particle. This affects the transparency of the results, which must be examined to filter out irrelevant unigrams and bigrams. Nevertheless, the results corroborate Aijmer's introspective work. Additionally, they offer evidence that there is an alternative method by which a reliable subset of translation equivalents can be semi-automatically determined. It also is a method that could be used to obtain a large set of disambiguated source language examples that can be investigated further to determine concrete source language internal clues for disambiguating particle functions.

2 Background

Discourse particles are often defined as uninflected lexical items that do not contribute to the propositional meaning of the utterances in which they occur. Their contribution is therefore often described as pragmatic, and said to express information about the speaker's attitudes or expectations (Stede & Schmitz 2000). It is also said to mark epistemic modality, to code speaker-hearer relations, to mark the evidentiality of the information or to mark discourse relations (Aijmer 1996). However, because pragmatic concepts are difficult to explicitly characterize, the exact contribution of a given discourse particle is often elusive. Additionally, many discourses particles are multifunctional, and a research aim is to determine how these functions can be reliably distinguished from each other. This work is compounded by the fact that different functions proposed for a particle are often closely related.

The Swedish particle *ju* is a prototypical discourse particle according to most definitions. It is uninflected and its removal from an utterance doesn't affect the truth-conditions of the utterance. *Ju* has only one other easily identifiable non-pragmatic usage which is easily distinguishable.[1] Aijmer (1977) was the first to discuss *Ju*'s pragmatic contribution to Swedish discourse. She originally identified four uses of *ju*, all having to do with speaker knowledge. The particle was argued to tend to occur in explanations (ex. (1)), in marking common knowledge (ex (2)), in utterances expressing criticism (ex. (3)) or in an utterence expressing the speaker's conviction (Examples from Aijmer, 1977).

(1) a. Du kan inte går ut. Det regnar **ju**.
b. You can't go out. It's raining.
(2) a. Kärnkraften gör **ju** att det inte blir någon arbetslöshet.
b. Atomic power will make it so that there isn't any unemployment.
(3) a. Du har **ju** inte kommit it tid en enda gång.
b. You haven't come in time a single time!

[1] Many discourse particles have homonyms that are adverbs or scalar particles, lexical items that do contribute to propositional meaning. These uses have to be distinguished from pragmatic uses. For *ju*, there is a semantic usage that appears in constructions of the form "*ju...desto*", and is roughly equivalent to "*the...the*" in "the more the merrier" or "the faster the better".

The first function is a case of *ju* being used to mark a coherence or rhetorical relation, while the other three all relate to the speaker's attitude towards the utterance.

Aijmer (1996) examined the translations in English of utterances in which the particle appear. This work was done with a small 10,000 word parallel corpus of Swedish and English. In addition to systematically giving the translations, another assumption of this work is that the different lexical items that occur in the translations of the particle in English may show systematic relationships with different functions of the particle. Thus the translations can be used as a method to determine particle functions.

In Aijmer's corpus, *ju* was one of the most frequent particles (310 examples out of the 775 found), but it is only expressed in the English translation in about 29% of the cases (89 cases). Aijmer (1996) identified four main functions: MODALITY, INTERACTIVE, INTERPERSONAL and DISCOURSE FUNCTIONS. As MODALITY Aijmer (1996) classifies cases where *ju* is translated as *I suppose* or *could*, expressing some degree of uncertainty as to the correctness of the information. An example from the Europarl corpus (Koehn, 2002) is given in (4). INTERACTIVE uses are identified as those examples that were translated by tag questions and *you know*, and are described as appealing to the hearer for agreement or acknowledgement of some information. INTERPERSONAL use is described as "emphasizing that the speaker and hearer have some knowledge in common, *ju* may create a feeling of intimacy and rapport" (1996:402), and an example from Aijmer is given in (5) (1996:421). The DISCOURSE FUNCTIONS are further divided into 3 categories, **emphasis** (translated by *just, only, surely, certainly,* clefts expressions, italic typeface), **expectation**, (*obviously, actually, of course, after all, as a matter of fact*) and **evidence**, (*since, as, because,* non-finite *-ing*). In its evidential usage, Aijmer argues that *ju* can "indicate that the hearer is appealed to as the source of knowledge" (1996:399). Examples of DISCOURSE usage taken from Aijmer (1995:422), (6) and the Europarl corpus, (7), are given below.

(4) a. … om alla de övriga medlemsstaterna accepterar det så kan en medlemsstat gå ur, och det kan man **ju** säga är en juridisk självklarhet.
b ... if all the other Member States agree, a Member State can terminate its membership, and that is, **I suppose**, self evident in law

(5) a. För dem har **ju** sig själva att uppfostra **ju**.
b. because they have their own selves to educate...

(6) a. ... och ja kan **ju** amerikanskan så jag tog platsen.
b. ... and **as** I know how to speak the American language I got the job.

(7) a. Det är ett nöje att se oss i ombytta roller, han skuggade **ju** mig för betänkandet om den övergripande strategin för reformerna av mänskliga resurser ....
b. It is a pleasure to see our roles reversed, **since** he shadowed me on the report on the overall strategy for the human resource ....

The MODALITY example illustrates a case where the particle doesn't correspond with any specific lexical item in English. The verb phrase "kan säga", literally "can say" could potentially be translated as *I suppose* even if the particle wasn't present, but the particle does seem to make clear that "kan" doesn't refer to ability but to uncertainty on the part of the speaker, i.e. uncertainty as to the correctness of the characterization of the ability of a Member State to terminate membership as being "self-evident in law".

The INTERPERSONAL uses seem to be cases where shared (private?) knowledge between the speaker and hearer is emphasized, many of the discourse functions translate into presupposition triggers in English, suggesting that the *ju* marked information is perceived as commonly accepted.

Among the DISCOURSE FUNCTIONS it is not clear how the different subtypes can be distinguished. The DISCOURSE **evidence** uses seem to be marking a typical reason relationship where the *ju* occurs in the discourse segment that expresses the reason why some other situation has obtained.

One of the difficulties in constructively applying Aijmer's results is that many of the translation equivalents identified occurred only once or twice in her data, which makes it difficult to determine if the translation is a common means in English by which to express the meaning, or if it is perhaps idiosyncratic to the particular example. Using a larger parallel corpus can help tell if these patterns are reliable, and then further introspective work can concentrate on studying the most consistent translation patterns which are probably good candidates either for translations or for strongly related co-occurring meanings.

3 Method

Two tests from statistical machine translation were used to study the particles. Both of these tests require a large amount of data to be reliable and therefore the Europarl corpus was used. Europarl (Release v1, Koehn, 2002) is an aligned, multilingual parallel corpus extracted from the proceedings of the parliament of the European Union and is freely available for 11 languages, including Swedish and English.[2] The corpus can be considered a mix of formal spoken language and read language in that much of the corpus consists of what seems to be short speeches by different speakers, but not in the form of a dialogue.

The Swedish corpus is 17,296,225 words (Koehn, 2002) while the English version of the corpus is 15,650,494 words (unix wordcount), resulting in or 626,771 aligned areas. From the Swedish corpus all aligned areas[3] containing *ju* and their English translation equivalents were excerpted. This resulted in 6022 aligned areas where *ju* occurred. All words and bigrams that appeared more than 20 times in the English translation areas were examined. χ^2 tests were used to identify unigrams and bigrams that were significantly more frequent in the translation area than in the rest of the corpus. These unigrams and bigrams are good candidates for translation equivalents (Manning & Schülze 1999). These words were then used as input to the ϕ^2 test.

The ϕ^2 test is a distance measure that determines the strength of the association between two variables.[4] It has therefore been used as a word alignment measure between a source language word and a possible target language translation (Church & Gale 1991; Gale

---

[2] Downloadable from http://www.isi.edu/ koehn/publications/europarl/

[3] Aligned areas are generally sentences pairs but in some cases there are 1-2, 2-1 and 2-2 sentence alignments.

[4] Let a = freq(source, target), i.e. the co-occurence frequency of the source word and target word, b = freq(source) – freq (source, target), c = freq(target) – freq (source, target) and N = all aligned areas, so d = N – a – b – c. Then :

$$\phi^2 = \frac{(ad - bc)^2}{(a + b)(a + c)(b + c)(b + d)}.$$

& Church 1991). A value 0 indicates no association and a value 1 indicates perfect association, thus the higher the value, the stronger the association. For each particle, ϕ^2 values were calculated for the source language particle and potential target language translations.

Finally, a number of examples of the most frequent co-occuring English unigrams and bigrams were examined manually to study how the results compared with the results in Aijmer (1996).

4 Results

4.1 Results of the χ^2 tests

382 unigrams and 705 bigrams were significantly more frequent in the areas aligned with occurrences of the particle. Those unigrams and bigrams that were likely translations of the meaning contributed by the particle are presented in Table 1. Words also found by Aijmer (1996) are given in the first column. The second column shows a selection of the other words identified as possible translation equivalents, and a few examples that could be translation equivalents that were however not identified by Aijmer (1996) are underlined.

On the whole, it seems that with this simple method a number of markers of the pragmatic information expressed by *ju* in Swedish, can be semi-automatically identified in English. However there are a number of drawbacks.

First, closer examination of actual translation pairs revealed that some of the translations do not seem to express the meaning contributed by the discourse particle in any way. This is not surprising, given that this meaning is often not an essential part of the message and c.f. Aijmer's (1996) results with only 29% of examples expressed. This introduces a certain amount of noise in the data.

Second, there are a large number of high-frequency words, in particular closed-classed words that are identified as significantly more frequent. This is a general effect of the way the language data interacts with the way the statistical tests are applied. Because χ^2 compares the frequency of unigrams and bigrams in aligned areas with each other, high-frequency words will have a tendency to occur in almost all translation areas with a high level of significance. These results have to be manually filtered out from the rest of the data.[5]

Some open class words were also significantly more frequent, but these words are certainly not translations of *ju*. For example, words like *Sweden, Denmark, beef* and *money* among others were more frequent in the translations. There are two possible explanations. This could be an effect of the distribution of the discourse particle in the Swedish originals. *Ju* occurs naturally in Swedish when the speaker is speaking Swedish. It is perhaps less likely that translators will add a discourse particle when translating from another language, as in most cases there won't be a corresponding expression in the source language. This would mean that a large number of the examples with *ju* were originally said in Swedish, rather than being a translation to Swedish from another language. This makes it understandable that certain other lexical items relating to topics of interest to Swedish

[5] The frequency at which a unigram or bigram becomes insensitive to these statistical measures could perhaps be determined by empirical testing, which could in turn automate this filtering process.

Words in Aijmer (1996)	**Sample of other words**			
actually	a	also	basic	cannot
after	able	always	basis	carried
all	about	amsterdam	be	case
as	above	an	become	central
because	absolutely	and	becomes	certainly
course	accession	another	becoming	clear
could	account	any	beef	clearly
fact	actual	anything	been	closer
just	administration	applies	behind	come
know	again	are	being	comes
matter	agenda	area	better	companies
obviously	ago	at	body	company
of	agreed	austria	border	could
since	allowed	available	borders	countries
surely	almost	away	but	country
you	already	bank	can	create

Bigrams in Aijmer (1996)	**Sample of bigrams**			
you know	a clear	a good	a matter	a problem
of course	a common	a great	a member	a question
after all	a country	a large	a more	a result
as a	a decision	a little	a new	a very
a matter	a European	a long	a number	a way
matter of	a few	a lot	a political	able to

Table 1 Words more frequent in the translations than the corpus as a whole that could be translations. All are significant to at least $p \leq 0.05$, and words in bold were also identified by Aijmer (1996).

parliamentary representatives would be more frequent here than in the corpus as a whole, e.g. discussion of Sweden or Sweden's neighbors.

Alternatively, these words may just be very high-frequency lexical items in this corpus, and thus their significance is just an effect of their general high-frequency and its interaction with the statistical measure.

In sum, χ^2 gives a rough picture of the meaning of a particle, but is not precise enough to allow any strong conclusions. Because a large number of unigrams and bigrams that aren't possible translation equivalents are also identified by the χ^2 measure, the results have to be filtered by hand, so a certain amount of subjectivity is introduced into the final results. The measure seems to be less successful with this data than with other lexical items. This is most likely an effect of the noise in the data, i.e. the particle often isn't translated in any identifiable way, and because of the diffuse or vague meaning that the particle, compared to other lexical items.

4.2 Results of the ϕ^2 test

The words with the top 35 ϕ^2 scores are shown in Table 2 along with a selected number of low-ranked unigrams. Immediately we can see that the ϕ^2 measure gives more specific, and

thus more useful information relating to the semantic contribution of the particle. Table 3 presents the same results for bigrams. We can observe at least one translation equivalent not identified by Aijmer (1996), *indeed*, which showed the 16th strongest relationship with the particle.

First we can see that a number of lexical items identified as translation equivalents in Aijmer's (1996) study appear among the highest scores, showing that they have the strongest relationship with the particle. These are underlined. We also find evidence of *ju*'s homonym in the comparative expressions that show a strong association. These are marked in italics.

Second, there are a number of vague, closed-classed terms that still appear to have a strong association. Again this is most likely an effect of applying the measure to this type of data. Very high-frequency words tend to show a high degree of association with almost any source language word.

Finally, we can see that unlikely translation equivalent open-class words such as *beef* and *Sweden* are ranked very low, showing a much weaker association. In this way they are distinguished from the closed-class high-frequency terms which still showed a high-level of association.

Thus with this measure it is possible to distinguish between lexical items that significantly co-occur with the particles, from those that co-occur as a side-effect of the context in which *ju* or *nog* is introduced into the text.

course	0.004200112315	which	0.0005341511216	since	0.0002485759374
all	0.001987486524	the	0.0004921076903	always	0.0002471448774
after	0.001903956728	but	0.0004345321883	…	
sooner	0.001853081119	have	0.0004298578296	*selected low-rank unigrams*	
is	0.001277616655	here	0.0003917315054		
because	0.001191055326	in	0.0003661550353	sweden	9.1772534048363e-05
we	0.000956710187	already	0.0003529397057	money	8.58572085727467e-05
fact	0.000934958332	as	0.0002876820396	denmark	4.5395313916179e-05
that	0.000748073560	then	0.0002859528643	beef	2.6363971871725e-05
it	0.000736206917	be	0.0002819315940		
know	0.000702382725	there	0.0002789101432		
more	0.000662376281	actually	0.0002749920182		
are	0.000652926089	*better*	0.0002744384420		
not	0.000628187501	also	0.0002591803255		
surely	0.000564424822	what	0.0002520769261		
indeed	0.000558607052	a	0.0002505986817		

Table 1 Top results of the ϕ^2 test for unigrams, as well as selected low-ranked lexical items.

after all	0.0139201456168712	*the longer*	0.00022212932663961
greater the	0.00638551854659638	is indeed	0.000210713826771468
of course	0.00512211275183667	is precisely	0.000205466332102801
the better	0.00386675075969963	has of	0.00020321663578044
more we	0.00333968610374183	as we	0.000197588023234919
the greater	0.0026718018828522	we have	0.000190803662797999
is after	0.00254104177740122	is that	0.000182143592718249
in fact	0.000924919149939707	we all	0.000170859028603137
more the	0.000922844468930272	course be	0.000150435576502129
it is	0.000382178255533762	is not	0.00014633947904675
we know	0.000359348765943205	know that	0.000146171184844111
you know	0.000339033539064875	case that	0.000143675587411756
that is	0.000293340549736507	is what	0.00012842803944562
fact is	0.000274597575005327	what we	0.000122115630824927
as you	0.000238195322027033	we cannot	0.000118543361783137
which is	0.000231196424477064	at all	0.000117098029632712

Table 2 Top results of the ϕ^2 test for bigrams.

4.3 Examples examined

A number of examples for each translation equivalent identified by Aijmer were examined further. Aijmer's four functional groups were not present to the same degree. For example, there were only three cases where *I suppose* was used to express *ju*, (an example was given in (4)) the MODALITY usage, and neither *suppose* nor *I suppose* were among the significant expressions identified by the χ^2 test.

Expressions listed as INTERACTIVE or INTERPERSONAL by Aijmer (1996) were also infrequent or even missing in the Europarl corpus. This is most likely an effect of the difference in corpus genre: Aijmer (1996) worked with a corpus of novels, including many children's stories, with dialogues between friends and family, while Europarl has only formal interactions between speakers in the parliament.

Instead, most translation co-occurrences were those words and bigrams that fell under the category of DISCOURSE FUNCTION. Recall that Aijmer (1996) distinguished between three subtypes: **emphasis**, **expectation** and **evidence**. Are these three subtypes distinct, and can we find subgroups of each function by examining the English translations? **Emphasis** is shown by the examples below with *just*, (8) and *surely*, (9).

(8) a. Det här är **ju** inte något svenskt förretag, utan vi diskuterar Europas framtid. b. This is not **just** some kind of Swedish project, we are discussing the future of Europe.

(9) a. Detta krav ä ju ändå förnuftigt !
b. **Surely** this is a reasonable demand ?

Utterances expressing **emphasis** did dominate the examples translated by *just* and *surely*. However, many of the examples also tended to support some claim, i.e. in (9), that the demand must in fact be reasonable is offered as support for granting the demand. Thus, many **emphasis** examples can also be analyzed as marking a claim, that is part of a

rhetorical relation, i.e. thus coinciding with an **evidence** function.

In the next set of examples, *ju*'s contribution is expressed in English using the translations listed as typical of the **expectation** function. Recall that for this usage the speaker is said to be marking the information as common knowledge, treating it as given, not questionable, possibly presupposing the information. Inspecting a number of examples suggests that translations with *in fact*, *the fact that*, or *indeed* express this type of meaning (*the fact that* and *indeed* not listed in Aijmer 1996).

(10) a. Det finns **ju** många olika sätt att skapa full sysselsättning på, men det är inte möjligt utan ekonomisk framgång och stark konkurrensförmåga.
b. There are, **of course**, many different ways to create full employment, but this cannot be achieved without economic progress and a high degree of competitiveness.

(11) a. Det är något nytt men det är egentligen inte så förvånande, det här var **ju** ett speciellt fall.
b. That is a new departure, but then again it comes as no surprise, for this was a special case **after all**.

(12) a. Herr talman ! Det är förvånande att följdrapporten om ansvarsfrihet den här gången tycks väcka mycket mer intresse än det egentliga beviljandet av ansvarsfrihet, som **ju** förbereddes av det gamla parlamentet.
b. Mr President, it is amazing that the follow - up report on the discharge seems to be generating far more interest this time than the discharge itself, which was, **in fact**, prepared by the previous Parliament.

Again, in example (11) and (12) also seem to be part of an **evidence** relation. In (11) the new departure is no surprise because it was a special case, and in (12) the speaker implies that the report should not have generated so much interest because it was prepared by the previous parliament.

The following examples illustrate clear cases of **evidence** relationships. The same translations identified by Aijmer (1996) seem to be reliable indicators of this usage. In addition, expressions such as own, (13) also seem to have the same function.

(13) a. Vad är det för mening med denna tjänst och denna befattning, när OLAF **ju** enligt kommissionens beslut ansvarar för behandlingen av alla svåra tjänstefel gentemot gemenskapens intressen,... ?
b.What purpose would this office and agency actually serve, when by the Commission's **own** decree, OLAF is responsible for all serious breaches of duty against the interests of the Community, ... ?

In (13) the claim that OLAF is responsible for serious breaches of duty is supported by the information that it was the commission itself that decreed as such, clearly an **evidence** relationship.

5 Discussion

5.1 Word alignment measures applied to discourse particles

χ^2 and ϕ^2 tests automatically present us with a number of co-occurring items that then have to be manually examined to find the relevant unigrams and bigrams. Thus the statistics don't give straightforward results, and this seems to be caused by the special characteristics of discourse particles.

Discourse particles differ in several ways from other lexical items with which these measures have been used. There are actually three different relationships that can be identified between the particle and a translation area. There are cases where there seems to be an identifiable expression or word that is the translation of the particle. For these examples, if the particle was removed this word or expression would be removed as well. These are close to true translation equivalents.

The second type of example is made up of cases where the particle consistently occurs in a certain type of context. The particle co-occurs with certain words, expressions or constructions, i.e. *ju* tends to occur in connection with a statement of belief or knowledge or with a conjecture, but these other expressions are certainly not the translation of the particle itself. Often the particle emphasizes or softens the semantic contribution of the other words or expressions in Swedish. In these cases the removal of the particle might result in a similar or even the same translation in English.

The third group of examples are those where there is no evidence of the particle in the translation, not even in the form of some systematic co-occurrence. These types of examples introduce noise into the calculation.

When χ^2 or ϕ^2 are applied to the first two examples we can often identify a number of translation co-occurrences, though the strength of the relationship may be less than with other lexical items. Actually, χ^2 and ϕ^2 tests cannot make any distinction between those unigrams and bigrams that are translation equivalents and those that are just co-occurring expressions, but this is actually positive for studying discourse particles. If we had a word-alignment measure that actually gave only exact translation equivalents, it would miss the co-occurrence relationships that are an important clue to the function of discourse particles.

5.2 Functional categories

From an examination of a small set of key translation equivalents it is apparent that it is difficult to distinguish the different DISCOURSE FUNCTIONS identified by Aijmer (1996) from each other. The **evidence** usage occurs in utterances that are rhetorically related to the previous discourse with a kind of reason relation. The other two uses, **emphasis** and **expectation,** do not characterize rhetorical relations that relate one proposition to another, but instead characterize two different types of epistemic modality.

The **evidence** relationships often co-occur with the other two relationships, which makes sense as all propositions in a coherent discourse have some sort of rhetorical relationship with the previous discourse, and such a relationship can certainly occur with an utterance that expresses some sort of epistemic modality.

There are also good explanations for why these types of propositional attitudes would co-occur with an utterance that expresses the reason why an earlier situation obtains, or why an earlier statement is true. Utterances expressing a reason relationship often co-

occur with emphatic statements because the speaker often strongly wants his or her statement to be accepted. Support for a claim or statement also tends to overlap with information that is marked as common knowledge or presuppositional, e.g. the **expectation** modality, because these types of information make strong arguments.

5.3 Future work

The next step is to study the examples for systematic clues to the different functions. Many different features that may play a role in categorizing particle usage, and these features have been discussed in other attempts at automatically disambiguating discourse particles. Scheler & Fischer (1997) looked at particle usage in task-oriented dialogues in German. They used clues such as turn-taking information, information about preceding and following syntactic phrases, the speaker's role, as well as the speech act of the preceding utterance as clues, to disambiguate using a connectionist network. The resulting network was able to correctly categorize about 50% of the particles. Working with the Verbmobile dialogues, Fischer & Pook (1998) use turn taking and the role of the utterance in the dialogue model as cues, and were able to automatically categorize 83% of the particles. Stede & Schmitz (2000) identify a number of features that they believe are useful in disambiguating discourse particles in German Verbmobile dialogues, including whether or not the particle is included in a collocation, its position in an utterance, syntactic features, such as the tense of the verb or if the particle occurs in the scope of a modal verb, as well as the previous dialogue, including the speech acts that preceded the use of the particle.

Many of the listed clues are only relevant for dialogue, and not the non-dialogue speech found in the European Parliament proceedings, but still, there are numerous features that could be examined. For example, it may be possible to distinguish **emphasis** uses from the others by length and the form as an exclamation, signaled by the exclamation mark used in (9), among other things. **Expectation** uses seem to occur often in contrastive statements. These utterances also tended to be shorter than other examples. The others functions relate closely to modality, and it is possible that features shown to be related to epistemic modality, such as the animacy or definiteness of the subject or what the tense of the phrase in which the particle occurs, might be useful features to investigate.

In conclusion, using simple statistical techniques for word alignment is a promising method to semi-automatically derive a subset of translation equivalents or co-occurring expressions for discourse particles, despite the special characteristics that discourse particles display. While these tests do not exhaustively identify all the means of expression, they easily generate a large number of examples of the most frequent uses, matching results found by manual work on a smaller amount of data. The results can also help point further research in the most fruitful directions.

Acknowledgments

The author gratefully acknowledges the Netherlands Organization for Scientific Research (NWO) for financial support, grant 355-70-005. All errors are my own.

References

Aijmer, L. (1977). Partiklarna *ju* och *väl*. *Nysvenska Studier*,[In Swedish].

Aijmer, K. (1996). Swedish modal particles in a contrastive perspective. *Language*

Sciences 18 (1-2), 393-427.
Church, K. and W. Gale (1991). Concordances for parallel text. In *Proceedigs of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*.pp 40-62, Oxford.
Fischer, L. and H. Brandt-Pook (1998). Automatically disambiguating discourse particles. In *Proceedings of Coling/ACL '98 Workshop on Discourse Relations and Discourse Markers*, Montreal.
Gale W. and K. Church. (1991). Identifying Word Correspondences in Parallel Texts. *Proc. DARPA NLP Workshop.*
Koehn, P. (2002). Europarl: A multilingual corpus for evaluation of machine translation. http://www.isi.edu/koehn/publications/europarl/.
Manning, C. and H. Schülze (1999). *Foundations of statistical natural language processing.* MIT Press, London, England.
Scheler, G. and K. Fischer (1997). The many functions of discourse particles: A computational model of pragmatic interpretation. in *Proceedings of CogSci.*
Stede, M. and B. Schmitz (2000). Discourse particles and discourse functions. *Machine Translation*, 125-147.