# Towards Automatic Identification of Discourse Markers in Dialogs: The Case of *Like*

**Sandrine Zufferey**
University of Geneva
School of Translation and Interpretation (ETI)
40, bd du Pont d'Arve
CH – 1211 Geneva, Switzerland
sandrine.zufferey@eti.unige.ch

**Andrei Popescu-Belis**
University of Geneva
ISSCO / TIM / ETI
40, bd du Pont d'Arve
CH – 1211 Geneva, Switzerland
andrei.popescu-belis@issco.unige.ch

## Abstract

This article discusses the detection of discourse markers (DM) in dialog transcriptions, by human annotators and by automated means. After a theoretical discussion of the definition of DMs and their relevance to natural language processing, we focus on the role of *like* as a DM. Results from experiments with human annotators show that detection of DMs is a difficult but reliable task, which requires prosodic information from soundtracks. Then, several types of features are defined for automatic disambiguation of *like*: collocations, part-of-speech tags and duration-based features. Decision-tree learning shows that for *like*, nearly 70% precision can be reached, with near 100% recall, mainly using collocation filters. Similar results hold for *well*, with about 91% precision at 100% recall.

## 1 Introduction

The identification of discourse markers (DMs) is an essential step in dialog understanding, since there is often a prosodic, syntactic and functional distinction between DMs and the rest of an utterance. For instance, the identification of DMs is relevant to lower-level analysis processes such as POS tagging or parsing.

After a brief theoretical definition in relation to natural language processing, this article will focus on the highly ambiguous discourse marker *like* – which besides a DM can also be a verb, a preposition, etc. As a DM, *like* mainly fulfills one function (to introduce an approximation with a variable scope), so the main problem in NLP is to disambiguate occurrences of *like* as a DM from other occurrences. We describe in section 5 two experiments that assess the performance of humans on this task in terms of inter-annotator agreement, then proceed to automate the identification of *like* as a DM, using collocation filters (section 6), a POS tagger (section 7), and decision-tree classification (section 8), which is also extended to the identification of *well*. The automated methods appear to be useful aids to manual annotators, since they reach 70% precision for *like* with near 100% recall.

This article is related to a dialog processing and retrieval application, developed within the IM2 project[1]. We make use of the ICSI Meeting Recording corpus of transcribed and annotated dialog, which contains 75 one-hour recordings of staff meetings, each involving up to eight speakers[2]. Each channel is manually transcribed and timed. We use here an initial release of 50 dialogs, annotated with dialog acts, segmented into about 65,000 prosodic utterances.

## 2 Role of Discourse Markers in Dialog

### 2.1 Definition

Despite the wide research interest raised by discourse markers for many years, there is no generally agreed upon definition of this term. The first difficulty arises from the fuzzy terminology used to designate these elements. Even though in English they are most often referred to as *discourse markers*, a variety of other names are also used, such as *discourse particles*, *discourse connectives*, *pragmatic markers*, etc. But the main problem for the study of DMs is that there seems to be no agreement regarding which elements should be included

[1] Interactive Multimodal Information Management, a project sponsored by the Swiss Government. This research is related to the Multimodal Dialogue Management module, see *http://www.issco.unige.ch/projects/im2/mdm*.

[2] See *http://www.icsi.berkeley.edu/Speech/mr/*. We are grateful to the ICSI-MR group for sharing the data as part of the IM2/ICSI agreement.

in this class. For instance, in English, Fraser (1990) has proposed a list of 32 DMs, but Schiffrin (1987) has only 23. Moreover, these two lists have only five common elements. The lack of agreement on what counts as a DM reflects the great diversity of approaches used to investigate them, resulting from divergent research interests, methods and goals.

At a very general level, it is nevertheless possible to formulate a rather consensual definition of DMs. Following Andersen (2001, p. 39), discourse markers are "a class of short, recurrent linguistic items that generally have little lexical import but serve significant pragmatic functions in conversation." Items typically featured in this class include (in English): *actually*, *and*, *but*, *I mean*, *like*, *so*, *you know*, and *well*.

Our study of DMs and its application to natural language processing is related to a wider-scope investigation of DMs which is grounded in relevance theory (Sperber & Wilson 1986/1995). In this framework, DMs encode a procedure whose role is to constrain the inferential part of communication, by restraining the number of hypotheses the hearer has to consider in order to understand the speaker's meaning[3].

### 2.2 Importance of Discourse Markers for NLP

The analysis of DMs for language processing is often inspired by discourse analysis theories such as Rhetorical Structure Theory (Mann & Thompson 1988). In this context, DMs are used to detect coherence relations automatically (Marcu 2000). For example, *so*, *therefore* and *then* are supposed to indicate a relation of *conclusion* between two segments. However, this analysis of DMs is not fine-grained enough: for instance, if the three markers above imply the same type of relation, why can they not be interchanged in every context?

More recently, DMs have also been used as useful cues to detect dialog acts and conversational moves. For example, *oh* implies a response to a new piece of information and *well* implies a correction (Heeman, Byron & Allen 1998). However, DMs are then only partial cues, since there is no one-to-one mapping between the use of a marker and the presence of a given relation (see for instance Taboada 2003).

In order to provide a more precise and comprehensive framework for the use of DMs in natural language processing, we derived elsewhere a three-step resolution procedure from a relevance-theoretic analysis (Zufferey 2004). These steps can be summarized as follows:

1. detect the occurrences of DMs
2. attach an inferential procedure to every marker
3. determine the scope of each procedure

In the remainder of this paper, we will focus only on the first step, i.e. the detection of DMs. The difficulty of this task comes from the fact that DMs are very ambiguous items. Typically, words like *well*, *now* or *like* can fulfill multiple functions. The first step towards a correct use of DMs for language processing is therefore to disambiguate them, i.e. to extract only the occurrences of the respective lexical item functioning as a DM – in other words the *pragmatic* occurrences (see their definition for *like* in section 3 below). Sections 6, 7 and 8 below will describe various automatic methods to accomplish this task. Note that even if we have grounded our approach in relevance theory, this first task is of paramount importance to any theory of discourse. For instance, in an RST framework, DMs can be used to infer coherence relations only if their pragmatic occurrences have previously been identified.

### 2.3 Overview of DM Frequencies

The manual annotation of DMs in a subpart of the ICSI meeting corpus (ca. 6 hours and 60,000 words) shows a big difference in the frequency of occurrence for various DMs. The most frequent ones are *but* (543 times), *like* (89), and *well* (287). Others are moderately frequent, e.g., *actually* (43), *basically* (21) or *now* (19), while other are very rare: *furthermore* (2), *however* (1), *moreover* (0). The frequency of each DM is relatively stable across the meetings.

The frequency of DMs depends a lot on the type of discourse. For example, the DM *however* is found much more frequently in written than in spoken language. There are about 50 occurrences of *however* in the London-Lund Corpus (500,000 words, transcription of spoken language) and about 550 occurrences in the Lancaster-Oslo/Bergen (LOB) corpus (1 million words, written texts). *However* – like most other DMs – is also much more frequent in dialogs as opposed to monologs. Another bias comes from the type of activity recorded: *however* is more frequent in formal settings, such as interviews *vs*. telephone conversations. And last, the regional variation of English, e.g. American *vs*. British, can influence the results. According to Lenk (1998), "*however* is not used in spoken American English".

The conclusion is that the frequencies above cannot be taken to be universal. But in the type of data we are interested in – dialogs – there is a high proportion of DM *like*. Besides, in a greater part of the ICSI-MR corpus (ca. 50 hours), 37% of the 2,116 occurrences of *like* correspond to its use as a DM. Hence the necessity to disambiguate it correctly becomes quite obvious, not only to have a better pragmatic analysis of occurrences but also to improve parsing and POS tagging[4].

[3] For a more detailed explanation of the role of DMs in relevance theory, we refer the interested reader to Blakemore (2002) for a recent survey.

[4] Sometimes, the POS tagging of a whole utterance can be ruined by an incorrect tagging of the DM (cf. section 7), not to mention its parsing.

## 3 The Case of *Like*

The discourse marker *like* is probably one of the most difficult to detect automatically because of the large number of functions of the word *like*. Apart from a DM, *like* can be used as a preposition, as in example (1) below, an adjective (2), a conjunction (3), an adverb (4), a noun (5) and a verb (6)[5]:

1. He was *like* a son to me.
2. Cooking, ironing and *like* chores.
3. Nobody can sing that song *like* he did.
4. It's nothing *like* as nice as their previous house!
5. Scenes of unrest the *like(s)* of which had never been seen before in the city.
6. I *like* chocolate very much.

The DM *like* is sometimes analyzed simply as a "filler", a hesitation word like *uhmm* that has no contribution to the meaning of an utterance[6]. However, other studies have shown that *like* has a much more complex role in dialogue. At a general level, *like* can be described as a "loose talk" marker (Andersen 2001). The function of *like* is to make explicit to the hearer that what follows the marker (for instance a noun phrase) is in fact a loose interpretation of the speaker's belief. Consider the following examples from the ICSI corpus:

1. It took ***like*** twenty minutes.
2. They had little carvings of ***like*** dead people on the walls or something.

In the first example, by using *like*, the speaker intends to communicate that the duration mentioned is an approximation. In the second example, the approximation concerns the expression that was used ("dead people"). By using *like*, the speaker informs the audience that this term doesn't exactly match what she has in mind. But *like* as a DM has also other functions, for example introducing a quotation (reported speech) and serving as a discourse link introducing a correction or a reformulation[7]. We will not elaborate on these functions, since the remainder of this paper will be dedicated to the identification of DM *like*, regardless of its precise functions.

## 4 Disambiguation of *Like* by Humans

Before trying to extract automatically the pragmatic occurrences of *like*, we have designed two experiments involving human judges. These preliminary experiments are useful indicators of the difficulty of this task, and the human scores will be used to assess more accurately the scores obtained by automatic methods systems.

[5] Adapted from the *Dictionnaire Hachette Oxford.* Oxford: OUP, 1994, 1943p.

[6] See for instance the Collins Cobuild English Language Dictionary (1987: 842).

[7] For a detailed analysis of *like*, see Andersen (2001).

### 4.1 Description of the Experiments

In the first experiment, human judges used only the written transcription of utterances containing *like*. In the second experiment, we explored the possibility to improve the level of inter-annotator agreement by using prosodic information: the human judges were also able to listen to the meeting recordings.

### 4.2 First Experiment: Annotation Based on Written Transcription Only

The first experiment involved 6 human judges, 3 men and 3 women whose age ranged from 25 to 40. They were divided in two groups of equal size: one of native English speakers, and one of French speakers with a very good knowledge of English.

Every judge was asked to annotate a number of utterances containing *like*, taken from two different sources: 26 occurrences came from the transcription of movie dialogs (from *Pretty Woman*) and 49 occurrences corresponded to one ICSI-MR meeting.

The participants were asked to decide for every occurrence of *like* whether it represented a DM or not. They were also asked to specify their degree of certainty on a three-point scale (1 = certain, 2 = reasonably sure, 3 = hesitating). Answers were simply written on paper. At the beginning, participants received written indications concerning the role of *like* as a DM as well as examples of pragmatic and non-pragmatic uses.

### 4.3 Second Experiment: Use of Prosodic Cues

In the second experiment, a group of 3 judges (2 French speakers and 1 English speaker) were asked to perform the same type of task, but in addition to the written transcription, they were also allowed to listen to the recording of the meeting when needed. This second experiment did not include dialogs from a movie but only from a one-hour ICSI-MR meeting, containing 55 occurrences of *like*[8]. The participants received the same set of instructions as in the first experiment, and in addition some explanation about the prosody of *like* as a DM. No time constraints were imposed, so the subjects could listen to the recording as many times as needed. On average, they completed the task in a half an hour. Access to the recording was provided through a hypertext transcript synchronized to the sound file at the utterance level (a multimedia solution developed for the IM2 project).

### 4.4 Results and Discussion

Results show that annotating DMs is a difficult task even for human judges. In the first experiment, the level

[8] Two of the participants had already participated in the first experiment, but the meeting was not the one used in the previous experiment.

of inter-annotator agreement measured by the Kappa coefficient is quite low ($\kappa = 0.40$) for the natural dialogs of the ICSI-MR corpus, and average for the movie transcription ($\kappa = 0.65$)[9]. In the second experiment, with the help of prosodic cues, inter-annotator agreement increases, and the annotation becomes much more reliable ($\kappa = 0.74$). Therefore, the identification of DM *like* is an empirically valid task, which can be accomplished at a reasonable performance level by untrained annotators. However, access to the prosodic information (from recordings) appears to be required. The inter-annotator agreement scores also set an initial boundary on automatic performances, which should not be expected to reach much higher levels. These results should be confirmed by experiments on longer transcripts, involving also annotators with specific training for DMs.

The results obtained in these experiments shed an interesting empirical light on a number of predictions that were made before the experiments.

First, it appears that DMs are easier to annotate in pre-planned dialogs, because such dialogs are less ambiguous than the natural ones. Indeed, the level of agreement reached for the movie transcription is much higher than for the ICSI-MR meeting in the same conditions (0.65 *vs.* 0.42). This result confirms that even if movie dialogs are made to reproduce the naturalness of naturally occurring dialogs, they are never as ambiguous, mainly because they only reflect the global communicative intention of one person (the author).

The second hypothesis we tested concerned the difference between native and non-native speakers' ability to annotate DMs. We believed that the group of native English speakers would have a better level of agreement. This prediction has not been confirmed: the group of non-native English speakers obtained nearly the same level of agreement as the native English speakers, for both types of corpora: $\kappa = 0.67$ *vs.* $\kappa = 0.63$ for the movie transcription and $\kappa = 0.4$ *vs.* $\kappa = 0.43$ for the meeting corpus. So it seems that non-native English speakers with a very good command of English are just as reliable as native English speakers to annotate DMs.

The third prediction we have tested is the possible correlation between the degree of certainty of annotators and the level of agreement. We haven't been able to find any significant correlation on both types of corpora and in both experiments. Thus, the capacity of human judges to evaluate their own intuition doesn't seem to be very high for this task. However, it should be mentioned that in general, the subjects have been much more confident in the second experiment, when they were able to use prosodic cues. The percentage of answers given with maximal certainty by the two annotators who took part in both experiments grew from 45% to 60% and from 65% to 87% respectively.

When looking more closely at the utterances upon which annotators do not agree, we can see that some types of occurrences of *like* seem to be much more difficult to annotate in both experiments. In most of these cases, *like* had the function of a preposition. For example, one subject was mistaken in annotating all occurrences of the type: *sounds like*, *seems like*, *feels like*, as DMs. This observation is not so surprising if we bear in mind that the pragmatic uses of *like* seem to have emerged (historically) in a grammaticalization process. According to Andersen (2001, p. 294): "the fundamental assumption here is that the pragmatic marker *like* originates in a lexical item, that is, a preposition with the inherent meaning 'similar to'". This suggests that more detailed explanations regarding the role of the DM *like* as well as some more training would probably improve the reliability of annotation.

To sum up, these two experiments have enabled us to quantify the level of agreement between human annotators and to confirm the usefulness of prosodic cues in order to efficiently detect the DM *like*.

## 5 Automatic Detection of *Like* as a DM

### 5.1 *A Priori* Cues

We have defined three linguistic criteria to be used for the disambiguation of DMs in general, which we will apply to the disambiguation of *like* in section 6 below.

The first criterion is the presence of collocations. For instance, when *well* is used to mark a change of topic, it is nearly always used in a cluster of markers such as: *well you know*, *well now*, *well I think* or *oh well*. On the contrary, when used to close a topic, *well* can very often be found in clusters like *OK well* or *well anyway/anyhow*. The criterion of collocations can also be applied the other way round, to establish cases where a given element *cannot* be a DM. For instance, when *like* is used in collocations such as: *I/you like*, *seems/feels like*, *just like*; or when *well* is used in constructions like: *very well*, *as well*, *quite well*, etc.

The second criterion is the *position in the utterance*. Again, depending on the word, this criterion can be used to ascertain that an element is a DM or, on the contrary, to rule out this possibility. For instance, w*ell* as a DM is nearly always placed at the beginning of an utterance or at least, at the beginning of a prosodic unit. In other cases, the use of this criterion implies that to be a DM, an element must not commence the utterance. According to Aijmer (2002, p. 30) : "Some of the discourse particles […] (*actually*, *sort of*) can, for instance, be inserted parenthetically or finally, often with little differ-

[9] We use Krippendorff's scale to assess intercoder agreement. This scale discounts any result with $\kappa < 0.67$, allows tentative conclusions when $0.67 < \kappa < 0.8$ and definite conclusions when $\kappa \geq 0.8$.

ence in meaning, after a sentence, clause, turn, tone unit as a post-end field constituent."

The third criterion is *prosody*. According to Schiffrin (1987, p. 328) "[a discourse particle] has to have a range of prosodic contours e.g. tonic stress and followed by a pause, phonological reduction".

However, even though these three criteria can help a human annotator to extract DMs successfully most of the time, some rare occurrences remain ambiguous. Some occurrences are at the boundary between a pragmatic and a non-pragmatic use. In these rare cases, both interpretations remain equally possible.

### 5.2 Application of *A Priori* Cues to NLP

Some of the criteria we propose seem relatively easy to automate. For instance, it is rather easy to extract a set a collocations once a list is made. Although some collocations imply the presence of a DM, and some other its absence, in some cases this criterion is in fact much more efficient in its second form, to rule out the presence of a DM. It is also rather easy to automate the criterion involving a certain position in the utterance, especially when the position is strongly constrained (for instance, at the beginning or end of the utterance). As far as prosody is concerned, the detection of pitch variations (for instance amounting to a correct transcription of commas) seems feasible for good quality recordings.

However, used independently from the others, none of these criteria can suffice to completely automate the extraction of DMs, even though in some cases a single criterion can be enough to get good results. For example, in the case of *well*, the position in the utterance can often be sufficient to correctly extract a significant proportion of all occurrences. Nevertheless, it will not solve all occurrences, since *well* is not always used at the beginning of an utterance but also at the beginning of a prosodic phrase, as in: "And I said, *well* I have to think about it". In these cases, the use of prosody to detect prosodic phrases becomes necessary. Similarly, the exclusion of some collocations like *very well*, *as well*, etc. is necessary to solve the last problematic cases.

In sum, these criteria seem to be sufficient to partially automate the disambiguation of DMs, which could serve to reduce the burden of human annotators.

### 5.3 Evaluation of NLP Performance

The evaluation of DM detection requires a "gold standard" (correct annotation) and the implementation of comparison metrics. The correct annotation of DMs was discussed in the experiments above, in the case of *like*, a highly versatile marker. In order to have enough data for our NLP experiment, one of the authors annotated manually all occurrences of *like* in 50 one-hour dialogs from the ICSI-MR corpus, generating 2,116 occurrences of *like*, of which 792 are DMs. About 20 occurrences of *like* could not be reliably disambiguated and were removed from the reference annotation.

We have already compared the annotations produced by human judges using the *kappa* metric. This metric can be used as well to score the performances of a system at distinguishing pragmatic from non pragmatic uses. Note that *kappa* compensates the scores by taking into account the probability of agreement by chance. A simpler but useful metric is the percentage of occurrences correctly identified, or accuracy. Unlike *kappa*, accuracy does not factor out agreement by chance, but provides a more interpretable score.[10]

Furthermore, if the task to be evaluated is the retrieval of pragmatic uses among all uses of the lexical item (which are trivial to detect), then recall and precision are also relevant. For instance, to evaluate techniques that filter out non-DMs, we will require them to reach nearly 100% recall, and a reasonable precision – say, more than 0.6 or 0.7 for *like*, i.e. twice the baseline precision, which is the frequency of the DM use.

## 6 Filters for the Disambiguation of *Like*

We first explore the possibility to use a list of collocations in order to identify occurrences of *like* as a DM in two different corpora, ICSI-MR and a transcription of Switchboard telephone conversations. The best use of this criterion is to maximize precision while keeping recall as close as possible to 100%, i.e. to rule out a maximal number of occurrences that are not pragmatic while keeping *all* the pragmatic ones. Such a partial identification can be used as a filter to reduce the number of occurrences that must be processed manually.

The list of collocations that exclude the presence of a DM contains for example collocations such as: *something like that*, *I like*, *looks like*, etc. The full list contains 26 collocations and was tested on two different corpora: first, on a subpart of the ICSI-MR corpus, with 6 hours of recording, and approximately 60,000 words; then on the Switchboard data, transcribed and annotated with DMs (Meteer 1995), with ca. 2,500 conversations and about 3 million words.

Our method reaches 0.75 precision with 100% recall on the ICSI-MR corpus, and 0.44 precision with 0.99 recall on Switchboard. The main goal of the filter is thus achieved: recall remains very high on both corpora. A precision of 0.75 for ICSI-MR means that a significant number of occurrences are correctly ruled out – the initial proportion of pragmatic uses is about 1/3, while af-

[10] Note that the probability of agreement by chance is here close to 0.5, given that 20–40% of the occurrences of *like* are DMs. When the proportion of DM occurrences is $\alpha$, the probability of agreement by chance is $(\alpha^2 + (1 - \alpha)^2)$, hence 0.68 for 20% and 0.52 for 40%.

ter the application of the filter it reaches 3/4, and none of the pragmatic uses was missed in the process.

The efficiency of the filter is smaller on the Switchboard data (0.44 precision *vs.* 0.75 for ICSI). In the ICSI-MR corpus, the precision obtained is probably the highest possible one with this filter, since the corpus was used as a development corpus, from which we have extracted our set of collocations. On the other hand, in the Switchboard corpus, the lower precision might also be due to the incoherent annotation. We used indeed the annotation of DMs that was already present in Switchboard, and this annotation is not entirely reliable. In fact, no real theoretical assumptions seem to underlie this annotation and according to Meteer (1995) the criterion to decide if an ambiguous case was a DM was "[…] if the speaker is a heavy discourse *like* user, count ambiguous cases as discourse markers, if not, assume they are not." In such circumstances, we can expect that the low precision of our system on Switchboard can at least be partly attributed to this lack of reliability.

Finally, our system has performed the same task as human judges in the first experiment (see section 4) on 49 occurrences of *like* in one ICSI-MR meeting. Interestingly, if we compare the average *kappa* obtained between humans and the kappa obtained between the system and all human judges, we get the same value ($\kappa$ = 0.42). Even though the results obtained by this preliminary system are quite tentative, this comparison with human judges seems to indicate that the performance is quite acceptable.

## 7 Use of a Part-of-speech Tagger

The use of a POS tagger for disambiguating pragmatic *vs.* non-pragmatic uses of *like* is a straightforward idea. Indeed, if the accuracy of the taggers on colloquial speech transcripts was very high, this would help filtering out many (if not all) of the non-pragmatic uses, such as cases when *like* is simply a verb.

We experimented using QTag, a freely available probabilistic POS tagger for English (Mason 2000)[11]. The tagger assigns one of the following tags to occurrences of *like*: preposition (IN, 1,412 occurrences), verb (VB, 509), subordinative conjunction (CS, 134), general adjective (JJ, 52), and general adverb (RB, 9).

These tags must then be interpreted in terms of DM uses. A simple attempt is to use the tagger as a filter, to remove verbal occurrences. Hence, a VB tag is interpreted as non-DM, and all the other tags as (possible) DMs. Unfortunately, the evaluation shows that such a filter is unreliable: recall is 0.77, precision is 0.38, accuracy 44%, and *kappa* is only 0.02, i.e. near random correlation. As expected, other interpretations of the tags do not lead to better overall results. The most significant figures are obtained when selecting only adjectival uses of *like* (tagged JJ) as potential DMs: the recall is of course very low, but precision is 0.74, which means that the JJ tag could be used as a cue for the presence of a DM use.

The main reason that explains the failure of the tagger to detect DM uses of *like* is that it was not trained on speech transcription, where *like* is quite frequent. A tagger trained on speech (supposing annotated data is available) could use some punctuation from the transcription to improve its accuracy, such as marks for interruptions and pauses that sometimes appear around DM uses of *like*. This could help it to avoid marking some of those occurrences as VB. A study by Heeman, Byron and Allen (1997) has shown that when specific tags are assigned to DMs and the tagging is done in the process of speech recognition, both the quality of tagging and the correct identification of DMs are significantly improved.

## 8 Statistical Training of DM Classifiers

The relevance of machine learning techniques to detect DMs and to improve manually-derived classification models has already been emphasized by Litman (1996). We have conducted machine learning experiments with the 2,116-occurrence data set, and confirmed the relevance of the filters defined in section 6 above, and the role of several additional features. The results obtained with *like* are also compared, at the end of this section, with an analysis on *well* as a DM.

### 8.1 Features for the Classification of *Like*

For each occurrence of *like*, we extracted the following features that we thought relevant to the DM/non-DM classification problem:

- presence of a collocation that rules out the occurrence as a DM; since *like* can be either the first word or the second word in the collocation, we separated this into two features;
- duration of the spoken word *like* computed from the timing provided with the ICSI-MR transcriptions, which was generated automatically;
- duration of the pause before *like*: 0 or more, or –1 if the utterance begins with *like* (the segmentation into prosodic utterances was also provided with the transcription);
- duration of the pause after *like*: 0 or more, or –1 if the utterance ends with *like*.

In order to classify each of the occurrences of *like* as either a DM or a non-DM, we used decision trees as

[11] QTag uses a variant of the Brown/UPenn tagsets, and was trained on a million-word subset of the BNC (written material): *http://web.bham.ac.uk/o.mason/software/tagger/*.

provided with the machine learning toolkit WEKA (Witten and Frank 2000)[12]. Since not all the features are discrete, we used the C4.5 decision tree learner (Quinlan 1993), or J48 in WEKA. For testing, we experimented both with separate training and test sets derived from the data (e.g. 1,500 *vs*. 616 instances), and by using 10-fold cross-validation of classifiers as provided by WEKA. Results being similar, we report below the latter scores.

### 8.2 Results for the Classification of *Like*

The best performance obtained by a C4.5 classifier is 0.95 recall and 0.68 precision for the DM occurrences, corresponding to 81% correctly classified instances and a *kappa* of 0.63. This is a significant performance, but it appears to be in the same range as the filter-based method (tested only on a smaller data set). And indeed, the classifier tree (see Figure 1 in the Appendix) exhibits as the first nodes the two classes of collocation filters defined *a priori* in section 6. This is a strong empirical proof of the relevance of these filters. Note that this criterion has not been used by Litman (1996) who focuses on a much more detailed analysis of the prosody along with some textual features.

Moreover, the next feature in the tree is the duration of the pause before *like* ('pause_avant'): it appears that a relatively long pause before *like* (greater than 240 ms) characterizes a DM in most remaining cases (70 out of 78). This matches our intuitions about the prosodic behaviour of *like* as a DM. The next features in the tree have quite a low precision, and may not generalize to other corpora. Tentatively, it appears that a very short *like* (shorter than 120 ms) is not a DM.

The best classifier tends to show that apart from the collocation filters, the other features do not play an important role. A classifier based only on the collocation filters achieves 0.96 recall and 0.67 precision for DM identification (80% correctly classified instances and $\kappa = 0.62$), which is only slightly below the best classifier. Is it that the time-based features are totally irrelevant? An experiment without the two collocation filters shows that temporal features *are* relevant: the best classifier achieves 67% correct classification, with $\kappa = 0.23$, that is, somewhat above chance. Again, among the first nodes of the tree are the interval before *like* and its duration (Figure 2 in Appendix). Also, a pause after *like* seems to signal a DM. Temporal features are therefore relevant to DM detection, but they are in reality correlated with collocation-based features, which supersede them when they can be detected.

The conclusions of this experiment with *like* are that the simple features designed until now, though particularly efficient given their simplicity, do not allow for more than 70% precision (at 100% recall) for the detection of *like* as a DM. Time-based features do not outperform collocation-based filters – though the former could generalize better to other DMs. This result is also particularly interesting considering the fact that human annotators performed significantly better when allowed to use sound files. The results suggest that prosodic features other than duration are relevant for the disambiguation of *like*. Further work on the prosody of *like* (e.g. pitch) should enable us to refine this criterion.

### 8.3 The Classification of *Well*

Using a similar procedure, we have applied C4.5 classification to the detection of *well* as a DM. On the same dialogs as above, we annotated the occurrences of *well* as a DM (579) among all occurrences of *well* (873). About 66% of all occurrences are DMs, which gives a baseline classification score (all occurrences considered to be DMs).

The features defined for *well* are similar to those used for *like*: collocation-based filters (with a different content) and time-based features. In addition, we defined a collocation-based feature that is supposed to ascertain the presence of a DM, namely collocations such as *oh well* or *OK well*. We also consider the occurrence of *well* at the end of an interrupted or abandoned utterance (ending on transcriptions by '= ='), a feature we hypothesize to indicate a DM.

The highest accuracy, 91% and $\kappa = 0.8$, is obtained by a classifier combining the collocation filters and the duration of the pause after *well* (cf. Figure 3 in the Appendix). This corresponds to 91% precision and 97% recall for the detection of DMs.

The use of the collocation-based filter alone – the one that rules out DM occurrences based on the previous word, e.g. *as well* – yields only slightly lower performance (90% with $\kappa = 0.79$). Again, this does not mean that all the other features are irrelevant. Rather, the time-based filter based on the duration of the pause after *well*, which includes the detection of *well* at the end of completed or interrupted utterances, produces a classification accuracy of 75% (and a low *kappa*, 0.45), with 77% precision and 96% recall on the identification of DMs only.

These results suggest that time-based features could generalize to a whole class of DMs, but for individual DMs, such features are outperformed by collocations filters based on patterns of occurrences. The definition of collocation filters for a set of DMs seems feasible, albeit somehow tedious.

[12] The Waikato Environment for Knowledge Analysis (WEKA) is made available by Ian H. Witten and Eibe Frank at *http://www.cs.waikato.ac.nz/ml/weka*.

## 9 Conclusion

This paper has presented several computational approaches to the disambiguation of discourse markers, with a focus on the highly ambiguous word *like*. Experiments regarding the human capacity to annotate reliably the discourse marker *like* show that relatively untrained annotators reach a *kappa* agreement of about 0.74, producing reliable, though not perfect, annotations – provided they have access to the sound files. Automatic performance of the identification task, using a set of collocation filters, can help annotators by discarding some of the non-pragmatic occurrences. However, POS taggers seem unable to disambiguate the occurrences of *like* in speech transcripts. Finally, the training of decision trees on about 2,100 occurrences of *like* confirms the relevance of collocation filters as the main features, followed by time-based features, while correctly classifying more than 80% of the occurrences of *like*, and more than 90% of those of *well*.

Future work should explore the relevance of other potential features. However, given the strong pragmatic function of DMs, it is unlikely that low-level features combined with machine learning will entirely solve the problem. As we have seen, POS tagging is quite unreliable on DMs, but POS tags from the surrounding words could serve as features for statistical training. More data and more reliable annotations will also help. Another promising approach is the generalization of classification features across several DMs, which will allow the detection of an entire class of discourse markers.

## References


Aijmer, K. *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins, 2002.

Andersen, G. *Pragmatic Markers of Sociolinguistic Variation: a Relevance-Theoretic Approach to the Language of Adolescents*. Amsterdam: John Benjamins, 2001.

Blakemore, D. *Meaning and Relevance: the Semantics and Pragmatics of Discourse Markers*. Cambridge: CUP, 2002.

Fraser, B. An Approach to Discourse Markers. *Journal of Pragmatics*. 1990, vol.14, pp. 383-395.

Heeman, P., Byron, D., Allen, J. *Identifying Discourse Markers in Spoken Dialog*. Proceedings of AAAI Spring Symposium on Applying Machine Learning and Discourse Processing. Stanford, 1998.

Lenk, U. *Marking Discourse Coherence: Functions of Discourse Markers in Spoken English*. Tübingen: Gunter Narr Verlag, 1998.

Litman, D. Cue Phrase Classification Using Machine Learning. *Journal of Artificial Intelligence Research*. 1996, vol.5, pp. 53-94.

Mann, W., Thompson, S. Rhetorical Structure Theory: Toward a Functional Theory of Text Organisation. *Text*. 1988, vol.8(3), pp. 243-281.

Marcu, D. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge: MIT Press, 2000.

Mason, O. *Programming for Corpus Linguistics: How to do Text Analysis in Java*. Edinburgh: Edinburgh University Press, 2000.

Meteer, M. *Dysfluency Annotation Stylebook for the Switchboard Corpus*. LDC, Working Paper. 1995, 28p. http://www.ldc.upenn.edu/Catalog/CatalogList/LDC99T42/DFLGUIDE.PS (06/17/2003).

Quinlan, J. R. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufman, 1993.

Schiffrin, D. *Discourse Markers*. Cambridge: CUP, 1987.

Sperber, D., Wilson, D. *Relevance: Communication and Cognition*. Oxford: Blackwell, 1986/1995.

Taboada, M. *Discourse Markers as Signals (or not) of Rhetorical Relations in Conversation*. Proceedings of the 8th International Pragmatics Conference. Toronto, 2003.

Witten, I., Frank, E. *Data Mining: Practical Machine Learning Tools with Java Implementations*. San Francisco: Morgan Kaufmann, 2000.

Zufferey, S. Une Analyse des Connecteurs Pragmatiques Fondée sur la Théorie de la Pertinence et son Application au TALN. *Cahiers de linguistique française*. 2004, vol.25, pp. 257-272.

## Appendix

The classifiers (C4.5 decision trees) built for *like* and for *well* that are reproduced here must be interpreted using the following rules. Starting with the root of the tree, occurrences of the respective DM (*like* or *well*) are classified according to the features appearing at the nodes (round shapes). Depending on the values of the features (branches of the tree), the occurrences are classified as a DM (1) or not (0). The rectangular boxes contain the class (0/1) and the number of instances correctly/incorrectly classified.

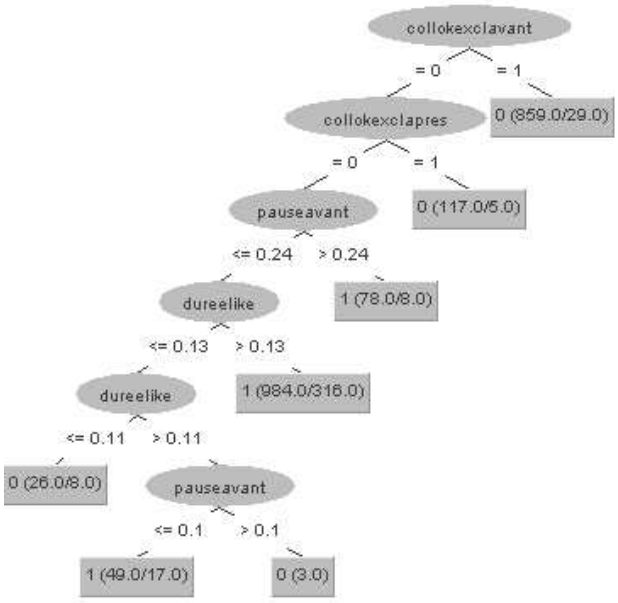


**Figure 1**. Best classifier for *like* as a DM.

In Figure 1, for *like*, the features can be glossed as follows: 'collokexclavant' – collocation filter ruling out the presence of a DM, depending on the word before *like*; 'collokexclapres' – collocation filter, word after *like*; 'pauseavant' – duration of the silent gap before *like*; 'dureelike' – duration of *like*. The most relevant feature, after the collocation filters, is the gap before *like*: a pause signals a DM in 91% of the cases.

In Figure 2, for *like*, when collocation filters are not used, a pause before ('pauseavant') and a pause after ('pauseapres') are the most reliable indicators of a DM (occurrences classified as '1').

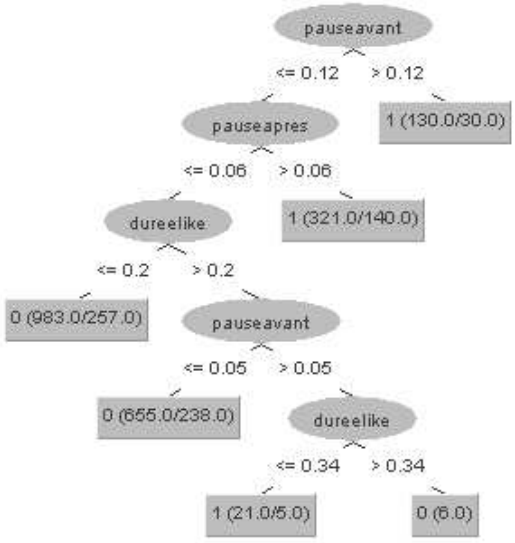


**Figure 2**. Best classifier for *like* as a DM, without the collocation filters.

In Figure 3, for *well*, the features are: 'collokexclavant' – collocation filter ruling out a DM, depending on the word before *well*; 'collokinclavant' – collocation filter that ascertains a DM, based on the word before *well*; 'pauseapres' – duration of the gap after *well*: –2 means that *well* is the last word of an interrupted utterance, and –1 means it is the last word of a completed utterance.

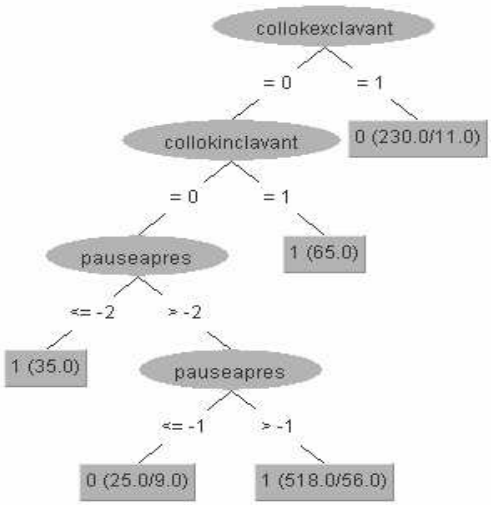


**Figure 3**. Best classifier for *well* as a DM.

The best classifier without the collocation features (not represented here) corresponds to the following rules: (a) if *well* ends an interrupted utterance, then it is a DM (100% accurate); (b) if it ends a completed utterance, then it is not a DM (88% acc.); (c) otherwise, it is a DM (81% acc.).