Automatic Identification of Discourse Markers in Multiparty Dialogues

Andrei Popescu-Belis[1] & Sandrine Zufferey[2]

ISSCO Working Paper 65

December 23rd, 2006

[1]ISSCO, School of Translation and Interpretation, University of Geneva. Bd. du Pont-d'Arve 40, 1211 Geneva 4, Switzerland. Email: andrei.popescu-belis@issco.unige.ch.

[2]Department of Linguistics, Faculty of Letters, University of Geneva. Rue de Candolle 2, 1211 Geneva 4, Switzerland. Email: sandrine.zufferey@lettres.unige.ch.

Abstract

The lexical items that can serve as discourse markers (DMs) are often multi-functional. *Like* and *well*, in particular, play numerous other roles apart from DMs: for instance, the first one can also be a verb and the second one an adverb. The goal of the present study is the identification, on transcripts of multi-party dialogues, of the occurrences of *like* and *well* that play a discourse or pragmatic role. DM identification is a binary classification task over the set of all occurrences of tokens *like* and *well*. The importance of DMs to computational linguistics is first discussed, along with previous experiments in DM identification. Then, the data is briefly described, emphasizing the DM annotation procedure and an inter-annotator agreement study. The proposed method uses lexical, prosodic/positional and sociolinguistic features, together with machine learning algorithms, among which decision trees are preferred. The results obtained using a ten-fold cross-validation procedure are analysed at length, focussing first on overall performance, and then on the relevance of each type of features. Feature analysis using a range of techniques shows that lexical indicators are the most reliable features for DM identification, followed by prosodic/positional features. Sociolinguistic features are slightly correlated with the use of *like* as DM, while the dialogue act of the utterance containing a DM candidate does not seem relevant to DM identification. A differentiated treatment for each token appears to improve performance in almost all experiments. The methods and features used here improve performance over the past experiments, and suggest that DM identification is a tractable problem provided enough training data is available for each DM type, and that lexical features are used for each type.

Keywords: discourse markers, *like*, *well*, automatic disambiguation, decision tree classifiers, machine learning, computational pragmatics.

Contents

Introduction

The identification of discourse markers is an essential step in dialogue understanding, since they are often prosodically, syntactically and functionally distinct from the rest of an utterance. Take for instance, the following utterance:

> This was like one of the first meetings I ever participated in.

This can have two quite distinct interpretations: "this *resembled* one of the first meetings..." or "this *was* roughly one of the first meetings...". The difference comes from the role of *like*, which acts as a discourse marker in the second interpretation. The speaker would probably give the token a marked prosodic contour—materialized by surrounding commas in writing—if the second interpretation was the intended one. The automatic identification of such occurrences is the overall goal of this article.

The notion of discourse marker (henceforth, DM) is briefly introduced from a linguistic and pragmatic perspective in Chapter 1, including a discussion of the sources of DM ambiguity for the lexical items *like* and *well*, which will be studied here. The DM identification task is defined and evaluation metrics to measure performance on this task are specified in Section 1.5. However, the outline of linguistic theories of the roles of DMs will be kept at a minimum, the focus being here the computational linguistics point of view.

The relevance of DMs to various computational studies of discourse is discussed in Chapter 2, along with an overview of previous methods that were proposed for DM disambiguation, and their results (Section 2.3).

We then profile the corpus that we used, the ICSI Meeting Corpus, in Chapter 3, focussing on the distribution of *like* and *well* for various speakers. In Chapter 4 we bring the empirical proof that DM disambiguation by humans is reliable, on condition that annotators can listen to the dialogue recordings during annotation. The failure to automatically disambiguate DMs using the results of a part-of-speech tagger is brielfy analyzed in Chapter 5.

A set of features that are relevant to DM identification is described in Chapter 6, including lexical collocations (Section 6.1), prosodic and positional information (Section 6.2), and sociolinguistic features pertaining to individual speakers (Section 6.3). These features are also studied in this chapter from a quantitative, descriptive point of view in relation to the DM vs. non-DM classification. The features can be used by a variety of classifiers constructed using machine learning, as shown in Chapter 7.

Automatic construction of decision trees using machine learning generates classifiers that reach disambiguation accuracies comparable to inter-annotator agreement scores, and well above various baseline scores. The best results also compare favourably to previously obtained ones. The results are extensively discussed in Chapter 8). The analysis of the relevance of various features to the DM identification problem indicates that lexical collocations are the most relevant indicators, followed by positional/prosodic features. The method can therefore be generalized to other DMs, provided enough training data is available.

Chapter 1

The Roles of Discourse Markers *Like* and *Well* in Human Dialogues

1.1 General Definition

Despite the wide research interest raised by discourse markers for many years, there is no generally agreed upon definition of this term and of its extension. A problem for the study of DMs is that there seems to be no agreement regarding which elements should be included in this class. For instance, in English, Fraser (1990) proposed a list of 32 DMs, but Schiffrin (1987) considered only 23. Moreover, these two lists have only five elements in common!

The lack of agreement on what counts as a DM reflects the great diversity of research interests, goals and methods. This diversity explains also why, along with the term 'discourse marker', a variety of other names are also used, such as discourse particles, discourse connectives, pragmatic markers, cue phrases, and so on. Authors like Fraser (1996a,b) use different terms to designate distinct sub-classes of elements. For instance, in Fraser's terminology, discourse markers form a grammatical category, and are a subclass of what he calls pragmatic markers. According to him, the latter category encompasses all non-truth-conditional linguistic elements; their role is to "signal the speaker's potential communicative intentions" (Fraser, 1996a, page 168).

At a general level, it is nevertheless possible to formulate a rather consensual definition of DMs. Following Andersen (2001, page 39), we will consider here that DMs 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* (often pronounced as *y'know*), and *well*.

A more specific definition can be provided within the framework of relevance theory (Sperber and Wilson, 1986/1995) by fleshing out the possible "pragmatic functions" mentioned above. According to this view, a lexical item playing the role of a DM encodes a procedure that constrains the inferential part of communication, by restraining the number of hypotheses the hearer has to consider in order to understand the speaker's meaning. More explanations regarding the role of DMs in relevance theory are available in Blakemore's 2002 survey (Blakemore, 2002).

1.2 Levels of Ambiguity of Discourse Markers

Lexical items serving as DMs can be ambiguous in several senses. First, these lexical items can play a discursive role in some contexts (also called pragmatic or DM role), and a non-discursive

role in other contexts (also called semantic or sentential role). This is quite a frequent ambiguity in English as well as in other Indo-European languages.

Second, even when acting as a DM, a lexical item can fulfil different pragmatic functions according to the context of use, for instance it could trigger different inferential procedures. Additionally, the scope of the DM, that is the span of speech or text that it applies to, can vary as well. Hence, the ambiguity of lexical items that are DM-candidates pertains to three different levels, a fact that requires also a three-step disambiguation procedure (see below)

There are several possible explanations for these ambiguities, but they are beyond the scope of this study. According to one possible line of thought (Hovy, 1995), discourse markers are employed to signal 'semantic interpropositional relations' as well as 'interpersonal intentions' in discourse, the number of which far exceeds the number of available markers in English, estimated at fewer than 1000. In addition, DMs also "signal the articulation of rhetorical structure", which is viewed as a surface discourse structure mapping more or less closely the interpropositional-semantic and interpersonal levels of relations. Therefore DMs often have both rhetorical and either interpropositional-semantic or interpersonal roles.

Another possible explanation of the DM/non-DM ambiguity of some lexical items is grounded in historical linguistics and grammaticalization theory. For instance, Traugott (1995) outlines a path of lexical evolution (a 'cline') that proceeds from clause-internal adverbial roles through sentence adverbial roles to discourse particle roles[1]. The three markers used to illustrate this path are *indeed*, *in fact* and *besides*, which have evolved from their literal meaning into potential DMs. Their main pragmatic (DM) function is roughly to indicate that the current utterance constitutes an elaboration of an earlier one (Traugott, 1995; Fraser, 1996a). However, these items still fulfil non-DM functions in specific contexts[2]. A large number of English DMs seem to retain their original meaning as well, though for some of them the DM function has been lexicalized and is far more frequent than the non-DM one.

The goal of the present study is to propose and discuss methods for automatically disambiguating potential DMs at the first level, that is, finding out for specific lexical items in a given context whether they fulfil a DM or a non-DM function, or in other words a pragmatic or a non-pragmatic role. We will use these terms throughout the paper in relation to our binary classification problem (DM vs. non-DM) and refer to the elements that are disambiguated simply as lexical items or candidate DMs. Indeed, the term 'discourse marker' or 'cue word' refers to a specific function of a lexical item rather than to the lexical item as a whole.

For comparative purposes, we focus on two lexical items, *like* and *well*, in an attempt to find out commonalities and differences among the most accurate DM disambiguation techniques, as well as the surface features that are most relevant to classifiers based on machine learning.

These two items are among the most ambiguous DM candidates. As shown in the following sections, *like* and *well* both have several DM and non-DM functions in dialogue, and they are both highly polysemous in their non-pragmatic roles, as they serve equally well as a noun, a verb or an adjective/adverb. *Like* and *well* are among the most frequently occuring DMs, and they have been often analyzed from various theoretical perspectives. For instance, Schourup (1985) focuses his study of "common discourse particles" on the three lexical items *like*, *well* and *y'know*. Leaving aside *y'know*, which has quite a different behavior due to its syntactic properties as an autonomous sentence, we will focus hereafter on *like* and *well*, starting with a typology of their possible pragmatic and non-pragmatic functions.

[1]The two best known clines (Hopper and Traugott, 2003) are the nominal cline—nominal adpositions become case markers—and the verbal cline—main verbs become tense/aspect/mood markers.

[2]For instance: "This is a book that has no basis in fact and serves no decent purpose" (adapted from Traugott, 1995, section 3.1).

1.3 The Case of *Like*

Like is probably one of the most ambiguous candidate-DMs because of the large number of possible functions of the token *like*, both in pragmatic and non-pragmatic roles. Therefore, *like* is also one of the most challenging candidate-DMs with respect to automatic disambiguation techniques.

When it is *not* used as a DM, *like* can be a preposition, as in the example (1) below, an adjective (2), a conjunction (3), an adverb (4), a noun (5) and a verb (6)[3]:

1. He was *like* a son to me.

2. He avoids 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. The TV broadcasted scenes of unrest, the *like(s)* of which had never been seen before in the city.

6. I *like* chocolate very much.

As a DM, *like* is sometimes considered simply as a filler, that is, an interjection or hesitation word such as *uhmm* that brings no real contribution to the interpretation of an utterance, as reflected for instance in one of the entries (n. 11) of the definition offered by a monolingual English dictionary, the Collins Cobuild Dictionary (Sinclair, 2001, page 898–899). Another pragmatic function noted in this general dictionary (n. 12) is the quotative one; both functions are considered to be informal and characteristic of spoken language and are accompanied by the note: "Some people do not like this use."

> 11 Some people say *like* when they are thinking about what to say next or because it has become their habit to say it. [...]
> 12 Some people say *like* when they are reporting what they or another person said, or what they thought about something.

Many studies in linguistics and pragmatics have shown that in reality *like* has a much more complex role in dialogue (Andersen, 2000; Siegel, 2002). At a general level, *like* can be described as a loose talk marker (Andersen, 2001). Its function is to make explicit to the hearer that what follows it (for instance a noun phrase) is in fact a loose interpretation of the speaker's belief. The approximation introduced by *like* always applies to the lexical items placed after it, and has a variable scope. More precisely, following Andersen (2001), it is possible to identify five different functions of the DM *like*, as illustrated in the examples below, extracted from the ICSI Meeting Corpus (see chapter 3), with added punctuation:

7. It took, *like*, twenty minutes.

8. They could either, *like*, make an arrow directly, or put a new node.

9. They had little carvings of, *like*, dead people on the walls or something.

10. He was *like*, yeah, I can make dogs raise their ears.

[3]These examples are adapted from a bilingual dictionary (Corréard and Grundy, 1994).

11. It might be that if you add a new thing pointing to a variable you just, *like*... it just overwrites everything.

In example (7), by using *like*, the speaker intends to communicate that the duration following the token (that is, twenty minutes) is an approximation, i.e. *like* is used here as a hedge. In (8), the DM *like* is used to indicate that the expression that follows it is an example. In (9), the approximation concerns the expression used (*dead people*) and not its content. In this case, *like* is a marker of metalinguistic use: by using it, the speaker informs the audience that this term does not exactly match what she has in mind. In example (10), *like* serves as a discourse link introducing a reformulation of one own's discourse.

Finally, in (10), *like* is used to introduce reported speech, that is, a quotation of someone else's or one own's past words. The quotative function of DM *like*, already mentioned in the dictionary definition above, is probably one of its most recent pragmatic functions, as it may have appeared only in the 1970s or 1980s, Schourup (1985) being one of the first linguists to notice it. This function is specifically analyzed as "a case of grammaticalization in process" by Romaine and Lange (1991), which accounts for some of the reasons of the ambiguity of *like* in spoken language. The quotative use, along with other pragmatic uses, are often associated to the so-called 'valley girl' language, a caricatural stereotype derived from the English spoken by younger generations in California.

At an even finer-grained level, some of the pragmatic functions of *like* can be sub-divided further on. For example, when *like* signals an approximation, this need not necessarily concern a numerical value as in example (7) above, but can also amount to a metaphor or a hyperbole, as shown below in examples (12) and (13) respectively (from the same corpus, with added punctuation):

12. You know, having the headset reminds me of, *like*, working at Burger King or something.

13. Because until then it was sort of *like*, everything was *like*, wonderful.

We will not elaborate on these functions, since the remainder of this paper is dedicated to the coarser-grained identification of the DM uses of *like*, as opposed to its non-DM uses.

1.4 The Case of *Well*

Although not as ambiguous as *like*, *well* can also fulfil a variety of pragmatic and non-pragmatic functions (see for instance Schourup, 2001). When it is not a DM, *well* can be an adjective, as in examples (14) and (15) below, an adverb as in (16) and (17), a noun (18), or a verb (19)[4].

14. It's as *well* not to offend her.

15. I do not feel very *well*.

16. He sings as *well* as he plays.

17. It rained all afternoon as *well*.

18. This book is a *well* of information.

19. Tears *well* up in her eyes.

[4]These examples are adapted from a bilingual dictionary (Duval and Sinclair Knight, 1995). Note however that no instance of *well* as a noun or as a verb was found in the ICSI Meeting Corpus (Janin et al., 2003) that we use in this study.

As a DM, the role of *well* is, abstractly speaking, to "signal that the context created by an utterance may not be the most relevant one for the interpretation of the next utterance" (Jucker, 1993, page 450)[5].

Indeed, during the course of a verbal exchange, the context evolves continually, though this movement is not always linear. For example, a speaker can make mistaken assumptions about his/her partner's current available context. It is precisely in these situations that the DM *well* is most likely to occur, as an indicator that the context needs to be reassessed.

Following Jucker (1993), we can identify four functions of the DM *well*, illustrated by the following examples from the ICSI Meeting Corpus (Janin et al., 2003) with added punctuation:

20. A: Is the rising pitch a feature, or is it gonna be in the same file?
B: *Well*, the rising pitch will never be hand-annotated.

21. A: We'd need to prune. Right? Throw things away.
B: *Well*, actually, you don't even need to do that with XML.

22. So they'll say, *well*, these are the things I want to do.

23. Oh, yes, but... *well*, uh, yes, but what I mean is that...

Example (20) represents the use of the DM *well* that is most commonly described by linguists. In this type of examples, *well* conveys some kind of insufficiency, showing that some of the background assumptions of one of the speakers are not entirely adequate. In example (20), speaker A's question implies that rising pitch will be annotated. For that reason, B cannot answer A's question without denying that assumption first. Example (21) illustrates the function of *well* at an interpersonal level as a 'face-threat mitigator', for example when a request has to be rejected or, when the speaker disagrees with her partner. In (22), the DM *well* introduces an instance of reported speech. Another function of *well* is to mark hesitation, as in (23).

This taxonomy of pragmatic effects of *well*—also known as 'procedural' or 'cognitive'—is also supported, for instance, by empirical evidence from a corpus of non-native English speakers (de Klerk, 2005). The four types of cognitive effects found by de Klerk (2005, page 1190) are characterized as: indication that the speaker needs time to contemplate (example (23) above); indication that the hearer should reconsider an assumption (example (20)); marker of discourse coherence (topic shift or narrative staging, close to example (22); marker of turn-taking or turn-giving (close to example (21))[6]. We will compare below in more detail the statistics provided by de Klerk (2005) to those obtained on the ICSI Meeting Corpus (see Section 3.3), in relation to the a priori observation of the frequencies of pragmatic vs. non-pragmatic uses of the two lexical items studied here.

1.5 Performance of DM Identification

1.5.1 Definition of the Task

To the levels of ambiguity outlined in Section 1.2 above correspond different disambiguation tasks. A complete disambiguation of lexical items serving potentially as DMs requires a three-step analysis procedure that we described elsewhere (Zufferey, 2004)

[5] Or, in slightly different terms, to "signif[y] that the most immediately accessible context is not the most relevant one for the interpretation of the impending utterance" (Jucker, 1993, page 435).

[6] As we will not distinguish the different pragmatic functions of *well* in the remainder of this study, this is the place to quote de Klerk's figures for the frequencies of the four types of pragmatic uses, which are, respectively, 64%, 23%, 11% and 2% (de Klerk, 2005, 1190). These figures are based on about 500 occurrences of *well* as a DM in utterances produced by non-native English speakers from South Africa.

- **Step 0:** identify the lexical items that could be DMs.
- **Step 1:** identify in each utterance the DM vs. non-DM occurrences of the tokens that can serve as DMs—the goal of the present study.
- **Step 2:** attach to each DM the correct inferential procedure chosen from the repertoire of possible procedures, identified *a priori*.
- **Step 3:** determine the scope of each procedure, that is, the elements of the utterance to which the procedure carried by the marker applies.

The automatic identification of the lexical items (step 0) is an easy task—at least for written texts or for accurate speech transcripts—as they show no variability in their form, at least in English. However, the identification of the occurrences that function as DMs (step 1) is a much harder tasks, but it is a task of paramount importance to any computational study of DMs in discourse, regardless of its theoretical stance. The other two disambiguation tasks (steps 2 and 3) are more specific problems that cannot precede DM identification. Therefore, given the current difficulties of automatic DM disambiguation (step 1), we will focus only on this task in the present study. Determining the scope of a DM (step 3) seems to be the most difficult step to automate, but remains nevertheless indispensable for an exhaustive description of the effects of DMs in discourse.

The identification of DM vs. non-DM uses of lexical item—here *like* and *well*—is a binary classification task over the entire set of occurrences of the lexical item. The remainder of this section will outline how to evaluate such a classification (done by a human or by a computer), i.e. how to measure its accuracy.

1.5.2 Evaluation Metrics

The evaluation of DM identification requires a gold standard (a ground truth annotation) and the implementation of metrics for comparing a candidate annotation (i.e. a hypothesized classification of a series of lexical items) with the gold standard. To produce the ground truth annotation, which indicates for each occurrence of *like* and *well* whether it functions as a DM or not, a number of experts compare their own judgments of each occurrence and agree on the correct classification, possibly eliminating the most ambiguous occurrences. In this study, as explained below in Sections 3 and 4, about 4300 instances of each lexical item were annotated on the ICSI Meeting Corpus (Janin et al., 2003).

To compare an annotation produced by an individual judge with the ground truth annotation, or for that matter the annotation produced by an automatic DM classifier with the ground truth, one has basically to count the number of occurrences of *like* (respectively *well*) which were assigned the correct category (DM or non-DM), and compare this number with the number of occurrences which were incorrectly classified. This simple evaluation metric is called *accuracy* or *percentage of correctly classified instances*. Such a score should be accompanied by an indication of the likelihood that the two annotators agree simply by chance on the right category, a probability that varies with the proportion of DM vs. non-DM uses of a lexical item[7]. Therefore, given the variability of this number, an idea is to incorporate this factor into a more eloquent evaluation metric.

[7]If the proportion of DM occurrences among all occurrences is α ($0 \leq \alpha \leq 1$), then the probability of agreement by chance is $\alpha^2 + (1 - \alpha)^2$. In the present case, the experts found that ca. 45% of the occurrences of *like* and ca. 88% of those of *well* are DMs, so the probabilities of agreement by chance when classifying them as DMs vs. non-DMs are, respectively, 0.51 and 0.79. The smallest value for binary classification tasks is 0.5; therefore, the probability of agreement by chance is nearly minimal for *like* and somewhat higher for *well*.

The *kappa* (κ) score (Krippendorff, 1980)—based on and sometimes confused with Cohen's α (Cohen, 1960)— was proposed in order to factor out the probability of agreement by chance. In theory, κ can be used to compare two or more annotations produced by human judges, or to score the performance of a system by comparing its output to a gold standard. However, since *kappa* appears to be sensitive even to small differences between annotations, its most frequent use has been for measuring inter-annotator agreement, which is in principle significantly higher than "agreement" between the gold standard and a system's output.

The values of κ vary from 1 for identical annotations, to -1 for totally contradictory annotations; $\kappa = 0$ means that there is no statistical correlation between two annotations. Krippendorff's scale for measuring inter-annotator agreement estimates that for $\kappa < 0.67$ the agreement is insufficient to allow any conclusion about the annotated phenomenon. For $0.67 \leq \kappa \leq 0.8$, the agreement allows tentative conclusions. However, only values of $\kappa \geq 0.8$ reflect a significant agreement between annotators, allowing definite conclusions about the reality of the phenomenon that was annotated—here, for instance, the DM vs. non-DM distinction. *Kappa* and its associated scale are widely to evaluate corpus annotation and automatic language processing tasks. A number of detailed analyses of the merits and limits of κ and of the above scale for are found in (Carletta, 1996; Di Eugenio and Glass, 2004; Craggs and McGee Wood, 2005).

Furthermore, given that the DM disambiguation task can be conceived of as the retrieval of DM uses among all uses of a given item, the *recall* and *precision* evaluation metrics are also relevant (originally used for information retrieval by VanRijsbergen, 1979). Recall is the number of correctly identified DMs among all existing DMs, while precision is the number of correctly identified DMs among all items identified as DMs by the classifier (be it human or automated). F-measure is the harmonic mean of recall and precision.

For instance, a method that reaches nearly 100% recall, and a reasonable precision, e.g. more than 70% for *like* (given a baseline precision of ca. 45%), could be used to filter out non-DMs as a prelude to human annotation of new, unseen data.

In the present study, all the above metrics will be used to assess the performance of human annotators, to evaluate automated DM identification techniques, and to find correlations of different features of the utterances with the DM / non-DM functions of *like* and *well*, which will help to characterize the behavior of these lexical items and to design more accurate classifiers.

1.5.3 Comparison with Previous Studies

For comparison purposes, it is important that the task is defined in similar terms and that the same evaluation metrics are used across studies. However, this is not always the case in the previous studies—overviewed in the next chapter (section 2.3)—which use the following metrics. All of them start with an accurate word transcription of the speech signal.

Hirschberg and Litman (1993) offer a detailed discussion of the relation of prosodic, part-of-speech and orthographic features to the DM vs. non-DM classification. These results involve mainly confusion matrices, i.e. 2×2 tables indicating DMs classified as DMs vs. non-DMs, and non-DMs classified as DMs vs. non-DMs. Since the authors are mainly interested in the accuracy of each factor with respect to each of the uses, no global performance metrics (such as accuracy or κ) are used in the study, though overall accuracy appears in the final table (Table 13).

Siegel and McKeown (1994) use only accuracy, i.e. the percentage of correctly classified instances, with scores reaching about 80%. Despite this high absolute value, we will show that taking into account the probability of correct disambiguation by chance (which is greater than 50%) greatly reduces the significance of such a score.

Litman (1994, 1996) uses as an evaluation metric the opposite of accuracy, i.e. the error rate of a classifier: the number of discourse (DM) uses classified as sentential (non-DM) uses, plus the number of non-DM uses classified as DM—i.e. the total number of misclassified tokens—in proportion of the total number of instances. As a percentage, the sum of correctly classified instances and error rate equals 1. In addition, Litman (1996, section 3.3) computes confidence intervals for the error rates of automatic classifiers using 10-fold cross validation (see section 7 below).

Heeman et al. (1998); Heeman and Allen (1999) attempt to combine utterance segmentation with part-of-speech (POS) tagging and DM identification (see below sections 2.2.1 and 2.3) on accurate transcripts of spoken dialogues. Therefore, the DM identification task amounts to the correct tagging of a closed class of lexical items (which is in fact only exemplified but never fully listed in their articles) with specific tags: AC for single word acknowledgments, CC_D for discourse conjuncts, RB_D for discourse adverbials, and UH_D for interjections (among which*well*). Two-word DMs are treated as two DMs if each of the words is a DM, and verbs used as DMs (e.g. *see*) are simply annotated as VB and therefore do not seem to be part of the DMs targeted by the study. The error metrics that are used are recall/precision for DM identification (seen as a retrieval task) and a modified error rate computed as the number of classification errors (DMs that were not tagged by one of the above tags and non-DMs unduly tagged) divided by the total number of correct DMs—and quite curiously not by the number of all tokens, since they do not count the non-DM uses of the ambiguous lexical tokens (Heeman, 1997, pages 58–60).

Chapter 2

Discourse Markers in Computational Linguistics

Discourse markers have been the object of numerous studies in computational linguistics, as they play a considerable role in a number of discourse processing tasks such as the construction of discourse structures or the identification of dialogue acts. Quite naturally, their use for a given task requires their identification in speech or written language, and a number of studies contain proposals for identification methods of increasing accuracy. The outcomes of the present study will be later compared with the most important antecedent proposals in the final discussion of the results of the DM disambiguation method proposed here (section 8.5.3).

The studies that are briefly discussed below do not consider a constant type of DMs, a fact that is also reflected in the terminology they use. Some studies focus on discourse connectives, a class of markers that function almost exclusively as links between two sentences of a discourse. Other studies consider cue words, which are a broader class of lexical items (or compounds) that convey specific discursive information, or, in other words, signal particular discourse functions. Given the variety of lexical items targeted under the name of discourse markers, their uses and the techniques to identify them vary considerably. For instance, the synthetic view of six studies from 1984–1986 offered by Hirschberg and Litman (1993, Table 14, pages 529–530) shows that the range of 'cue words' analyzed by these studies varies considerably, as do the meanings hypothesized for them. Therefore, not all the studies discussed below are directly comparable, in goal and methods, to our attempt to disambiguate occurrences of *like* and *well*[1].

The first two sections of this chapter provide some examples of the *interest* for the fields of computational linguistics and natural language processing of the identification (or disambiguation) of DMs in spoken or written language. The two main motivations for the study of DMs from a language processing perspective are therefore the relation of DMs to discourse structure and their impact on other, more local discourse phenomena[2]. These perspectives will be overviewed respectively in sections 2.1 and 2.2 below. Quite naturally, all of the studies in

[1]The meanings assigned to these two DMs by the studies quoted by Hirschberg and Litman (1993, Table 14) represent a subset of those described in the previous chapter. These are, for *like*: support, comparison, example, repair, restriction; and for *well*: repair, response. Most of these meanings come in fact from Schiffrin's 1987 study quoted by Hirschberg and Litman (1993).

[2]This is summarized for instance by Heeman and Allen (1999, page 4): "DMs are conjectured to give the hearer information about the discourse structure, and so aid the hearer in understanding how the new speech or text relates to what was previously said and for resolving anaphoric references (Hirschberg and Litman, 1993). Although discourse markers, such as *firstly* and *moreover*, are not commonly used in spoken dialogue Brown and Yule (1983), a lot of other markers are employed. These markers are used to achieve a variety of effects: such as signal an acknowledgment or acceptance, hold a turn, stall for time, signal a speech repair, or signal an interruption in the discourse structure or the return from one."

this section presuppose a component that identifies DMs (or disambiguates candidate DMs) though in most cases the component is not precisely defined, as noted also by Heeman and Allen (1999, section 7.3, page 40): "Although numerous researchers have noted the importance of discourse markers in determining discourse structure, there has not been a lot of work in actually identifying them." The lesser priority of the DM ambiguity problem is probably due to the fact that the DMs under study, which are often discourse connectives, are not so highly ambiguous as *like* or *well*. The solutions proposed specifically for DM identification occupy therefore a smaller range of studies, which will be outlined in section refsec:other-dm-id below.

2.1 DMs as Primary Indicators of Discourse Structure

Many studies hypothesize the existence of specific discourse structures that organize the sentences or the utterances of a discourse. The studies often envisage DMs as a means of signalling (from the speaker's point of view) and respectively reconstructing (from the hearer's point of view) the discourse structure of a given discourse. Although some approaches use the term 'discourse connectives' to refer to this type of DMs, others prefer to use the more generic term 'cue words', or even 'clue words'—which is why it is important to summarize below the main tenets of this type of approaches. With respect to the two DMs that are the target of the present study, one should note that *like* is almost never categorized as a connective, while *well* sometimes is.

Many researchers have clearly noted the contribution of DMs to inferring discourse structure from linguistic form and context information. For instance, Hirschberg and Nakatani (1996) summarize the various knowledge sources for discourse processing as follows:

> Discourse structural information can be inferred from orthographic cues in text, such as paragraphing and punctuation; from linguistic cues in text or speech such as cue phrases (also called discourse markers or discourse particles, items that explicitly mark discourse structure) and other lexical cues; from variation in referring expressions, tense and aspect; from knowledge of the domain, especially for task-oriented discourses; and from speaker intentions (adapted from Hirschberg and Nakatani, 1996, page 286)[3].

In the excerpt above, Hirschberg and Nakatani (1996) quote the following studies that use cue phrases to infer discourse structure: Cohen (1984), Reichman (1985), Grosz and Sidner (1986), as well as two studies by Passonneau and Litman, later subsumed into a longer article (Passonneau and Litman, 1997)[4]. This latter study deals with discourse segmentation into ut-

[3]The exact quote reads: "Previous research has found that discourse structural information can be inferred from orthographic cues in text, such as paragraphing and punctuation; from linguistic cues in text or speech such as CUE PHRASES [*footnote:* 'Also called DISCOURSE MARKERS or DISCOURSE PARTICLES, these are items such as *now*, *first* and *by the way*, which explicitly mark discourse structure.'] (Cohen (1984); Reichman (1985); Grosz and Sidner (1986); [...]) and other lexical cues [...]; from variation in referring expressions [...], tense and aspect [...]; from knowledge of the domain, especially for task-oriented discourses [...]; and from speaker intentions [...]." (Hirschberg and Nakatani, 1996, page 286).

[4]Passonneau and Litman (1997, section 2.2) make in fact the same analysis as (Hirschberg and Nakatani, 1996) of the cues used to infer discourse structure: "The segmental structure of discourse has been claimed to constrain and be constrained by disparate phenomena, e.g., cue phrases (Hirschberg and Litman, 1993; Grosz and Sidner, 1986; Reichman, 1985; Cohen, 1984), plans and intentions (Carberry, 1990; Litman and Allen, 1990; Grosz and Sidner, 1986), prosody (Hirschberg and Pierrehumbert, 1986; Chafe, 1980a; Butterworth, 1980), reference (Webber, 1991; Grosz and Sidner, 1986; Linde, 1979), and tense (Webber, 1988; Hwang and Schubert, 1992; Song and Cohen, 1991). However, just as with the early proposals regarding segmentation, many of these proposals are based on fairly informal studies."

terances, with no hierarchy being constructed, and therefore we will separate it in our discussion below from the more hierarchically-structured approaches, with which we will start.

2.1.1 Cohen (1984)

Robin Cohen's (1984) study of argumentative dialogues introduces a tree-based representation of discourse structure which relates propositions according to their role in the argumentation. Discourse connectives play a central role as clue words for the (re)construction of this structure by the hearer, or for that matter by an automatic analysis algorithm. In fact, the types of relations between propositions correspond to categories of discourse connectives, to which specific structural positions are associated. For instance, a sentence S_2 following S_1 and containing a clue word of the type 'detail' such as *in particular* will be connected as a son of S_1 in the tree representation of the discourse. If S_2 contained a clue word of the type 'parallel' such as *in addition*, it would have appeared as a brother of S_1.

The concept of clue word considered by Cohen, including also multi-word expressions, is clearly larger than the notion of a discourse marker or a connective, as shown by the sample list of clue words categorized according to their function (Cohen, 1984, Appendix III, page 257). The clue words "coinciding with the connective taxonomy" are divided into five categories: parallel (e.g. *first, then, moreover, last*), summary (e.g. *thus, in conclusion*), reformulation (e.g. *in other words*), detail (e.g. *for example*), inference and contrast. In addition, other clue words are attitudinal expressions indicating a degree of belief, or emphasis expressions defending a claim, or constitute transitions.

The study acknowledges the fact that such clue words are not always obligatory in discourse, and can be supplemented or replaced by other types of "indirect" discourse-structuring devices such as lexical equivalence, reference or ellipsis. The relation to contemporary work on discourse focus and anaphora is briefly discussed. However, the study does not indicate methods to determine whether a candidate expression plays or not a clue word role in context.

2.1.2 Reichman (1985)

Reichman's (1985) book is one of the first attempts to construct a computationally tractable theory of discourse, which is implemented into a model based on ATNs (augmented transition networks). Similarly to Cohen (1984), this linguistic theory of discourse grants a significant role to clue words (or markers) as signals of relations between context spaces—the major device for discourse modelling in this theory. In Reichman's own words:

> Besides the specific expectations set up by the initiation of particular conversational moves, there is an additional mechanism that discourse participants use in the development and identification of context space structures. Speakers use specific surface linguistic signals--*clue words*—that usually accompany different types of conversational moves in a discourse. [Note:] We don't distinguish between cases where clue words are obligatory or not (Reichman, 1985, page 36)).

A similar observation was also made by Jurafsky and Martin (2000, sections 19.7-19.8), in a passage on cue words in NLP: "Algorithms for inferring intentional structure in dialogue work similarly to algorithms for inferring dialogue acts. [...] Machine-learning algorithms rely on features like cue words and phrases (Reichman, 1985; Grosz and Sidner, 1986; Hirschberg and Litman, 1993) or prosody (..., Hirschberg and Nakatani, 1996), and other cues. [...] Segmentation algorithms use boundary cues such as: (1) cue words like *well*, *and*, *so*, that tend to occur at beginnings and ends of utterances (Reichman, 1985; Hirschberg and Litman, 1993), (2) specific word or POS sequences (n-grams) that often indicate boundaries."

As for the identification and processing of clue words, the study provides a list of examples of conversational moves and "some of the clue words associated with them", as well as the subsequent evolution of context spaces following the specific conversational moves. In the list appear moves such as support (signalled by an initial *because* or *like*), interruption (*by the way*), or return to previously interrupted context space (*anyway*, *in any case*). The capacity for "conversational move and clue word classification to direct generation and interpretation of subsequent utterances" appears thus clearly among the "features that any grammar (processor) of discourse should possess" (Reichman, 1985, page 188), but the identification of the correct move from ambiguous (or worse, absent) clue words is not discussed any further.

2.1.3 Grosz and Sidner (1986)

The role of DMs as indicators of discourse structure was also described by Grosz and Sidner (1986). The authors consider that "certain words and phrases and more subtle cues such as intonation or changes in tense and aspect" are "among the primary indicators of discourse segment boundaries" (Grosz and Sidner, 1986, page 177). Given the particulars of this model of task-oriented dialogue, the functions of cue phrases can either "indicate changes in the intentional structure or in the attentional state of the discourse" (*ibid.*). Cue phrases can thus convey information about attentional change (corresponding to push/pop to/complete the focus stack), about the relation of the current intentions to previous ones, and about the precedence relationships between discourse segments' purposes—such as 'satisfaction-precedence' (*first, second, moreover*) and 'new dominanc' (*for example*) (section 6 and figure 12, pages 196–198).

As was the case with the previous studies, the authors acknowledge that "the cases listed here do not exhaust the changes in focus spaces ... nor have we furnished a set of rules that specify when cue phrases are necessary" (*id.*, page 199). The ambiguity of some of the most common markers is also mentioned, e.g. for *first, second, moreover* which can signal satisfaction-precedence but also new dominance, or for *now* and *next* can be used to signal, beyond a change of attentional state, either the creation of a new focus space or the return to a previous one. Finally, the precise identification and use of DMs may be related to the more general problem of recognizing utterance-level intentions, mentioned in conclusion as an open problem.

2.1.4 Mann and Thompson (1988)

Another influential theory of discourse structure is the Rhetorical Structure Theory or RST (Mann and Thompson, 1988). With respect to the previous approach, rhetorical relationships "provide a metalevel description of discourse" although they "could be recast as a combination of domain-specific information, and general relations between propositions and actions [...] and between intentions" (Grosz and Sidner, 1986, page 202).

Nevertheless, Marcu (2000) has successfully designed an automatic discourse parser—mainly applied to planned written text—based on RST, and making use of DMs among other indicators of rhetorical relations between discourse segments. For example, for example, *so*, *therefore* and *then* are assumed to indicate a relation of 'conclusion' between two segments.

While the efficiency of this type of cue for discourse parsing was demonstrated, an analysis of DMs based exclusively on RST seems insufficient to account for the specificity of each marker. For instance, in the previous example, although the three tokens signal the same relation, they are far from being interchangeable in every context. In a recent study of the correlation between rhetorical relations and DMs, Taboada (2006) has shown that there is no clear one-to-one mapping between the use of a DM and the presence of a given rhetorical relation, and that a majority of relations were not signals by DMs. DMs can therefore serve only as (very) partial cues to detecting rhetorical structure.

2.1.5 The Penn Discourse Treebank

An alternative approach is adopted for research on the annotation of discourse relations in the Penn Discourse Treebank (PDTB)—e.g. (Prasad et al., 2004). Contrary to RST-based studies, this approach does not start from abstract rhetorical relations, but still makes use of DMs to identify the structure of discourse and to hypothesize relations. According to the proponents: "low-level discourse structure and semantics [...] result (in part) from composing elementary predicate-argument relations whose predicates come mainly from discourse connectives and whose arguments come from units of discourse" (Prasad et al., 2004, page 88).

The PDTB annotation of discourse connectives (Miltsakaki et al., 2004) focuses on the relations between clauses that are signalled by subordinate or coordinate conjunctions, discourse adverbials, and even implicit conjunctions. Markers that are specific to spoken dialogue, such as those studied here, are not annotated. The PDTB focuses therefore on discourse connectives, and does not seem to consider the problem of ambiguous tokens (DM vs. not DM).

2.2 DMs as Cues for the Detection of Other Discourse Elements

DMs have also been used in computational approaches to derive other types of discourse information, pertaining to a somewhat lower structural level than the previously discussed approaches. These include the segmentation of spoken discourse into utterances, the recognition of the dialogue acts carried by utterances, the improvement of speech recognition, and the automatic categorization of discourse connectives.

2.2.1 DMs as Cues for Utterance Segmentation

Cue words that include DMs have been used as potential indicators of utterance or discourse unit boundaries in a number of studies. As noted by Heeman and Allen (1999, page 5), "DMs tend to be used at utterance boundaries, and hence have strong interactions with intonational phrasing. In fact, Hirschberg and Litman (1993) found that discourse markers tend to occur at the beginning of intonational phrases, while sentential usages tend to occur mid-phrase." This type of information has been used for utterance segmentation by Stolcke and Shriberg (1996); Passonneau and Litman (1997); Heeman and Allen (1999) among many others.

For instance, using similar DMs as Hirschberg and Litman (1993), Passonneau and Litman (1997) include cue words and cue phrases into their algorithms for speech segmentation into utterances, along with other linguistic features based on the presence of referential noun phrases. These features are shown to relevant to automatic segmentation methods, which perform at a comparable (albeit lower) level of performance as human annotators. DMs are used as cue word features for segmentation in the following way:

> The cue phrase features are also obtained by automatic analysis of the transcripts. Cue_1 is assigned 'true' if the first lexical item in P_{i+1} is a member of the set of cue words summarized in (Hirschberg and Litman, 1993)[5]. $Word_1$ is assigned this lexical item if cue_1 is true, 'NA' (not applicable) otherwise. Cue_2 is assigned 'true' if cue_1 is true and the second lexical item is also a cue word[6]. $Word_2$ is assigned the second lexical item if cue_2 is true, 'NA' otherwise. As with the pause features, the cue phrase features were motivated by previous results in the literature. Initial phrase position (cue_1) was correlated with discourse signaling uses of cue words in

[5]That is: *also, and, anyway, basically, because, but, finally, first, like, meanwhile, no, now, oh, okay, only, or, see, so, then, well, where.*

[6]That is, one of the following words: *and, anyway, because, boy, but, now, okay, or, right, so, still, then.*

(Hirschberg and Litman, 1993). A potential correlation between discourse signaling uses of cue words and adjacency patterns between cue words (cue_2) was also suggested. Finally, Litman (1994) found that treating cue phrases individually rather than as a class ($word_1$, $word_2$) enhanced the results of Hirschberg and Litman (1993).

2.2.2 DMs as Cues for Dialogue Act Recognition

The utility of DMs for the detection of dialogue acts or of conversational moves—as suggested already by Reichman (1985)—has been investigated as well.

In the context of the TRAINS spoken task-oriented dialogues, Byron and Heeman (1997) analyze the correlation between DMs and conversational moves, and between DMs and speech acts in adjacency pairs, though they do not design a dialogue act tagger. (The results of this study appear also in (Heeman et al., 1998)). The study focuses only on utterances starting with one of the four following DMs, *and, so, well, oh*, as they are the only ones that occur frequently enough in initial position to allow some conclusions. The problem of the potential ambiguity of *and* and *well* (discourse vs. sentential) is thus simplified as initial occurrences of these tokens in speech are almost certainly DMs.

The study shows that initial *and* is predominantly used to elaborate a plan, initial *oh* to respond to new information, and *so* to conclude—the number of occurrences of each DM being respectively 28, 17 and 28. There are also 7 instances of initial *well* which are all associated to a correction move. Regarding correlation with speech acts, the only significant results show that *and* and *so* correlate almost exclusively with acknowledgments (second parts in adjacency pairs).

Samuel 1999 also uses DMs as cues to dialogue acts, among other types of expressions. The method he uses to assess the relevance of cue phrases for dialogue act tagging somehow blurs the distinction between DMs and other particles. Samuel et al. 1999 also attempt to detect cue phrases based on several methods, in order to use them as indicators for dialogue act tagging. Their definition of cue phrases corresponds to a slightly more general notion than DMs, and the identification method they propose makes use (in the learning phase) of the predictive power of the cues for dialogue act recognition, followed by lexical filtering of the candidates. The resulting impact on dialogue act tagging compares favourably to various other methods based on cue phrases, such as using all the phrases, using manually-selected ones, or those with high mutual information figures. However, it seems that this method has no fine-tuned strategy to recognize only DM uses of ambiguous tokens.

2.2.3 DM Recognition to Increase the Accuracy of Speech Recognition

A number of studies have tried to increase the accuracy of automatic speech recognition by constraining recognizers with information from the discourse level. For instance, Stolcke et al. (2000) showed improvement of speech recognition accuracy when information about the dialogue acts carried by utterances was made available to the recognizer.

DMs have also been used as a knowledge source for speech recognition by (Heeman and Allen, 1999), in an enriched statistical language model that includes POS tags, DMs, speech repairs and intonational phrases. Therefore, the identification of DMs becomes part of a joint task, in which it aids (and is aided by) POS tagging and speech repairing. The authors claim that by solving these tasks simultaneously, they "obtain better results on each task than addressing them separately" (Heeman et al., 1998, page 1). In fact, the use of DMs reduces the number of POS errors from 1,219 to 1,189, a 2.5% reduction[7]. This error rate reduction holds, according

[7] "Although the improvements in perplexity and POS tagging are small, they indicate that there are interac-

to the authors, if confusions between discourse and sentential uses of DMs are not counted in the overall accuracy of POS tagging—probably because if they were, the number of POS would increase due to the difficulty of disambiguating discourse vs. sentential uses, as shown in chapter 5 below. It appears indeed that general purpose taggers seem to be unable to handle DMs. Sometimes, the POS tagging of a whole utterance can be ruined by an incorrect tagging of the DM (not to mention its parsing).

The fact that certain lexical items function as DMs is captured in this study by specific part-of-speech tags will be discussed in the section 2.3.4 below focussing on the techniques for DM identification.

2.2.4 Recognition of the Polarity of DMs

Recent studies by Ben Hutchinson (2004a; 2004b) target partially the problem of DM identification, and partially the problem of automatic categorization of some aspects of their discourse function, which is non-trivial for complex, multi-word discourse connectives.

The corpus of DMs collected from the Web by Hutchinson (2004b) contains 140 types of discourse connectives, as defined in Knott's (1996) classification. The collection method selected only non-ambiguous DM uses, with ca. 90% accuracy, using a parser. Therefore, examples of non-DM uses or uncertain occurrences of the tokens do not appear frequently. The method requires only filters with high precision, since recall is irrelevant given the large size of the Web. The resulting corpus consists mainly of sentences from written texts, with an average of ca. 30,000 tokens per discourse connective. The corpus does not target *like* or *well*, which do not seem to function as connectives.

The DM disambiguation task defined by Hutchinson Hutchinson (2004a) on the corpus described above aims at classifying tokens that are *known* to be DMs, according to three dimensions: polarity (positive or negative), veridicality (yes or no) and type (causal, temporal or additive). For most connectors of English, including phrasal ones, these values are known in advance, which provides the ground truth—as the method attempts to find these automatically. The sentences are parsed and a set of features is defined: lexical co-occurrences (frequent words, plus main/subordinate clause information), position, embedding, nature of verbs in clause, pairing of features of main and subordinate clauses. Several machine learning methods are tested, and the best scores reach 10% error rate.

The applications of Hutchinson's disambiguation method differ radically from ours. The classification of the tokens is already known in English, although Hutchinson points out that only 150 connectives from Knott's 1996 list of 350 have been properly classified yet. However, it seems that a manual categorization of these connectives, which must be done just once, could be sufficient. In contrast to our present study, Hutchinson does not target the ambiguity of connectors, though this is acknowledged for "and", which can sometimes have negative polarity, but should be classified as positive (Hutchinson, 2004a, page 686). The theoretical stance adopted in our study is, for ambiguous DMs or connectives, that each *occurrence* should be assigned its specific function. Hutchinson's corpus could be used for such a task, but requires manual identification of ambiguous DMs.

2.3 Methods for the Identification of DMs

In this section, we summarize the main proposals and findings of four studies that have explicitely focussed on the problem of DM identification, i.e. separating sentential from discourse

tions, and hence DMs should be resolved at the same time as POS tagging and speech recognition word prediction (Heeman and Allen, 1999, page 18).

uses of a number of candidate DMs, regardless of the further use of DMs for other CL/NLP tasks. The data used or produced by these studies, along with the procedure for hand-annotating the ground truth status of candidate DMs, are discussed in more detail in section 4.3 below.

2.3.1 Hirschberg and Litman (1993)

Hirschberg and Litman (1993) have conducted one of the first studies that dealt specifically with the problem of identifying DM vs. non DM uses of a number of lexical items. The authors note that the ambiguity problem and its computational solution had not received proper treatment before their study: "The question of cue phrase sense ambiguity has been noted in , although only cursory attention has been paid to how disambiguation might take place. A common assumption in the computational literature is that hearers can use surfae position within a sentence or a clause to distinguish discourse from sentential uses. In fact, most sytems that recognize or generate cue phrases assuume a cononical (usually first) position for discourse cue phrases within the clause" (Hirschberg and Litman, 1993, pages 504–505).

Taking into consideration the insufficiency of such assumptions to dismbiguate candidate DMs in spoken language, the authors proceed to define a model for DM disambiguation based on prosodic or intonational information in American English, drawing inspiration from pilot studies of *now* and *well* in a radio show recording. These studies illustrate the interest of studying individual DMs, although the goal is to infer a more general model; these studies are also replicated by Litman (1996).

The overall DM identification model generalizes the observations made on *now* and *well*. It consists of two sets of rules or 'models' to identify discourse uses, and two models to identify sentential uses. For instance, the 'discourse B' model indicates that a token that appears in the initial position of a multi-word phrase, and is deaccented or has a low intonational tone, is likely a DM. The authors proceeds to test the generality of the model by applying it to a set of 34 types of single word DMs[8], corresponding to 878 classifiable tokens—see section 4.3 below for more information about the data. For each candidate DM, the prosodic features required by the models were manually extracted prior to the application of the model.

In addition, the authors also propose, based on empirical evidence, that among the transcript-based features, the "presence or absence of preceding punctuation and part of speech [were] most successful in distinguishing discourse from sentential uses" (Hirschberg and Litman, 1993, page 523). A simpler model based only on 'orthographic' cues—comas, periods, dashes and paragraph breaks—was also tested.

The prosodic model correctly classifies 75.4% of the 878 classifiable tokens (i.e. those agreed upon by the two human judges), and 85.3% of the 495 tokens that were not coordinate conjunctions, which appear to be more difficult to disambiguate (31 types, excluding *and*, *or* and *but*). The confusion matrixes provided by the authors (tables 6 and 7) allowed Heeman (1997, pages 58–60) (see also Heeman and Allen (1999, section 9.3, page 40)) to compute recall and precision for the task of DM retrieval only, which amount respectively to 63.1% and 88.3%. Using the same formulae, recall and precision for DM retrieval among non conjuncts are respectively 82.7% and 81.5%. In addition, it is also possible to compute from the same tables the value of κ agreement between the human judges and the intonational model: we find $\kappa = 0.52$ for the whole set of classifiable tokens and $\kappa = 0.69$ for non-conjuncts.

The transcript-based model correctly classifies 80.3% of the tokens and the orthographic model alone correctly classifies 80.1% of the tokens. Using Table 11 from the article, it appears

[8]These are: *actually, also, although, and, basically, because, but, essentially, except, finally, first, further, generally, however, indeed, like, look, next, no, now, ok, or, otherwise, right, say, second, see, similarly, since, so, then, therefore, well, yes.* Quite surprisingly, *anyway* does neither appear in this list, nor in the list of non considered tokens discussed by Hirschberg and Litman (1993, page 517).

that recall and precision for the orthographic model amount respectively to 57.3% and 82.6%, and $\kappa = 0.54$. The scores of the orthographic model appear to be higher than those of the intonational model, which is probably due to the fact that the orthographic model is based on high-level information from human transcribers, namely the correct transcript of all words and the correct punctuation[9].

Hirschberg and Litman (1993) thus prove that both types of models, intonational and transcript-based, are relevant to DM identification. However, both models rely on substantial external information from human annotators, namely prosodic contour and transcript. The challenge of future study is thus to increase identification accuracy while decreasing the amount of required external knowledge.

2.3.2 Siegel and McKeown (1994)

Siegel and McKeown (1994) propose a transcription-based method for DM identification, which uses decision tree classifiers that are constructed thanks to a genetic algorithm. Unlike the hand-coded models proposed by Hirschberg and Litman (1993), the decision trees are optimized by the genetic algorithm to increase disambiguation accuracy. However, similarly to the orthographic methods proposed by Hirschberg and Litman (1993), the features used for the decision trees are the previous and following tokens with respect to the candidate DM and the type of the candidate DM itself. More precisely, the utility of tokens situated 2, 3 or 4 words after the candidate was tested as well, but was never selected by the learning algorithm. To increase avoid over-training, only neighbouring words that appear 15 times or more in corpus were kept as features, along with punctuation marks[10]—the use of punctuations presupposes that a considerable amount of knowledge is embedded in the transcript.

The same 34 types of DMs as Hirschberg and Litman (1993) are targeted by Siegel and McKeown (1994), and the training/test data is also the same, except that a larger chunk of the monologue is used, totalling 1,027 tokens. The authors only use accuracy as a performance measure, and do not provide confusion matrixes which would allow one to compute recall, precision and κ, therefore their results are more difficult to compare with others.

One of the baseline scores is the accuracy of a binary decision tree inspired from the orthographic model proposed by Hirschberg and Litman (1993) which can be paraphrased as: if the candidate DM is preceded by a comma or a period (i.e. if it is utterance-initial) then it is a DM, otherwise it is not. This binary tree reaches 79.16% accuracy. Unfortunately, after a 58-fold cross-validation training/test procedure, the average score of the best (and more complex) decision trees found by the genetic algorithm is onl 79.20%. Although the overall performance does not seem to be improved, the decision trees that were constructed by the genetic algorithm display a number of interesting linguistic rules, which were found automatically (Siegel and McKeown, 1994, Tables 3 and 4). For instance, constructions such as *the like* or *as well* indicate that *like* and *well* are used sententially, and not as DMs (we will extend this type of findings in section XXX below). The article also "demonstrates the utility of allowing decision trees to discriminate between clue words", as decision trees exhibit different tests for different tokens.

[9]Similarly, question marks entered by human transcribers are very reliable indications of questions (for a dialogue act labelling task), but are difficult to assign automatically.

[10]The full set of lexical features is thus: *a, and, are, as, at, can, for, I, in, is, it, of, that, the, this, to, we, you* along with the period, comma and apostrophe punctuations. If the neighbouring token of a candidate DM is neither of these, it is labeled 'default'.

2.3.3 Litman (1996)

The relevance of machine learning techniques to DM identification was further emphasized by Litman (1996) in a set of experiments that extended and completed earlier studies (Hirschberg and Litman, 1993; Litman, 1996) by improving manually-derived classification models. Most of the experiments targeted the same set of 34 DM types as above, over the same data set as Hirschberg and Litman (1993), with the same annotation of DM vs. not DM uses performed by the two authors and discarding the tokens disagreed upon.

The machine learning algorithms include the C4.5 decision tree learner that we will also use in this study, and the CGRENDEL algorithm that constructs sets of conditional rules. The resulting classification models are logically equivalent, and although their performances appear to be comparable, CGRENDEL rulesets are more often used in the study. The input to these methods are training examples from a set of candidate DMs (90% of the data) presented as sets of (feature, value) pairs for each candidate DM; the accuracy of the classifiers is then tested in the remaining data (10%). Repeating this operation ten times with different subsets of candidate DMs leads to 10-fold cross-validation scores.

The features used for classsification have a crucial influence on classification accuracy. Following Hirschberg and Litman (1993), Litman (1996) uses prosodic features assigned by human annotators, textual features extracted from human transcripts (including correct punctuation), part of speech information (assigned automatically), and the nature of the token (candidate DM) itself. The prosodic features are based on a theory of English prosody Pierrehumbert (1980) and extend the features used by Hirschberg and Litman (1993)[11]. The textual features include mainly indicators of the preceding and succeeding punctuations and cue phrases, which are sometimes duplicated into more or less abstract features, to test which one will be used by the classifier learner[12].

The part of speech is assigned to the candidate DM by an automatic tagger, which may be responsbile for some errors (see chapter 5 below). Curiously, neither the identity nor the part of speech of neighbouring words is used as features, although Siegel and McKeown (1994) had shown that they can be relevant to DM disambiguation. It could be hypothesized that these features showed too much variability given that the study targeted 34 DM types but had only about 900 examples available.

Litman (1996) provides a wealth of comparative results in various experimental conditions, as well as observations regarding the most relevant features for DM identification. However, results are difficult to synthesize from so many conditions. A constant observation is that most of the prosodic and textual models that were learned automatically outperformed corresponding models defined *a priori* by humans.

The best performance using all available features is 16.9% error rate (i.e. 83.1% accuracy) on the whole set of 878 occurrences, and 16.6% on the 495 non-conjuncts (with respectively, confidence intervals of ± 3.4 and ± 4.1 computed using 10-fold cross-validation). The article does not provide confusion matrixes, so it is not possible to compute recall, precision and κ.

However, it turns out that some of the classifiers constructed using less features outperform

[11] The prosodic features are the length and the position of the candidate DM in the intonational and intermediate phrases, the composition of the intermediate phrase (i.e. containing only the candidate, containing also other cue words, or other) and the prosodic accent in detailed or abstracted form (i.e. with more or fewer possible values) (Litman, 1996, Figure 2, page 61). Following Pierrehumbert (1980), "a well-formed *intonational phrase* consists of one or more intermediate phrases followed by a boundary tone. A well-formed *intermediate phrase* has one or more pitch accents followed by a phrase acccent" (Litman, 1996, page 57).

[12] For instance, 'preceding orthography' (i.e. punctuation) can be one of the following: 'comma', 'dash', 'period', 'paragraph', 'false' (i.e. none), 'NA' (i.e. not available, for 39 candidates who were recorded but not transcribed). The more abstract version, 'preceding orthography*' considers as possible values 'true' (i.e. present), 'false' (i.e. absent) and 'NA' (Litman, 1996, pages 61–62).

the general classifier, a phenomenon which is due to the learning procedure (because theoretically, using more features can only improve classification, otherwise the extra features can be discarded). The best overall results are obtained by a classifier constructed using the phrase-related prosodic features (i.e. excluding accent) and the identity of the candidate DM—this classifier has 14.5%±3.3 error rate (i.e. 85.5% accuracy), and 12.6%±3.3 on the non conjuncts[13]. The results appear thus to vary considerably across feature sets, and therefore it is not easy to identify the most useful features, though in general the candidate's part of speech and information about *suceeding* words or punctuations do not seem to increase performance. The availability of the 'token' feature, which allows the classifier to process DMs differently according to their nature, does not improve the performance of classifiers using all available features, but improves the performance of many models that use partial prosodic and textual features. This result, which confirms the same observation made by Siegel and McKeown (1994), shows that DMs have different behaviors and they should be processed specifically.

Finally, an experiment that allowed classifiers to mark candidate DMs as ambiguous, using a larger set of candidates (including the 75 ones upon which Hirschberg and Litman (1993) disagreed), increased the error rate of the best classifiers (those that used all available features) to 22.4%±4.1. As the authord indicates, "learning how to classify cue phrases as *unknown* is a difficult problem [which needs] more training data [or] additional features" (Litman, 1996, page 86).

2.3.4 Heeman and Allen (1999)

The problem of DM identification was coupled to speech recognition, utterance segmentation, POS tagging and above all repair detection and correction, applied to the Trains corpus (Heeman, 1997; Heeman et al., 1998; Heeman and Allen, 1999). As regards DMs, the study adopts a point of view oriented towards recognition of DMs rather than disambiguation of DM candidates, which is reflected in the corpus statistics that are provided and the performance measures that are used. The overall goal of the paper is to increase the accuracy of speech recognition by using a language model that incorporates knowledge about parts of speech, DMs, repairs, etc. The paper focuses on the language model itself, which is evaluated against manual transcripts of the dialogues using six fold cross-validation, in terms of perplexity decrease and POS/DM error reduction. The POS and DM probability distributions are estimated by training on annotated data, using a decision tree and a richer history of the previous tags, but no acoustic cues apart from pauses in speech.

DM identification is coupled with POS tagging, so that DMs receive specific tags:

> "DM usage is captured by the POS tags. The tag AC marks single word acknowledgments, such as *okay*, *right*, *mm-hm*, and *no*. The tag CC_D marks discourse conjuncts, such as *and*, *so*, and *but*. The tag RB_D marks discourse adverbials, such

[13]Heeman and Allen (1999, section 9.3, page 40) indicate that "th[is] algorithm achieved a success rate of 85.5%, which translates into a DM error rate of 37.3%, in comparison to the rate of 45.3% for Hirschberg and Litman (1993)." These figures are, however, unaccurate. The error rate is defined as the number of erroneous classifications (i.e. DMs classified as non DMs and vice-versa) over the total number of tokens, or occurrences of candidate DMs. Conversely, success rate or accuracy is the number of correct classifications over the total number of ambiguous tokens, so that the two sum up at 100%. However, Heeman (1997, pages 59–60) does not take into account the total number of tokens, but only the number of DMs, which increases the error rate artificially (since the denominator of the fraction is smaller) and could lead potentially to error rates greater than 1. Here is how Litman's "error rate" of 37.3% was computed by Heeman: "we can compute our standardized error rate by first computing the number of tokens that were incorrectly guessed: $14.5\% \times 878 = 127.3$ [*sic*]. We then normalize this by the number of DMs, which is 341. Hence, their error rate for DMs is $127.3/341 = 37.3\%$." Our suggestion is that, if a perspective based only on DM retrieval is adopted, recall and precision should be enough to characterize performance, with no need for a modified error rate.

as *then, now, actually, first*, and *anyway*. Finally, UH_D marks interjections, such as *oh, well, hm*, and *mm*. Verbs used as DMs, such as *wait*, and *see*, are not given special markers, but are annotated as VB. No attempt has been made at analyzing multi-word DMs, such as *by the way* and *you know*; however, phrases such as *oh really* and *and then* are treated as two individual DMs" Heeman and Allen (1999, page 9).

Unfortunately, we could find no complete list of the candidate DMs considered in these studies (Heeman, 1997; Heeman et al., 1998; Heeman and Allen, 1999). Observations from the previous quote and from the 'AC' classification tree (Heeman et al., 1998, Figure 2) suggest that at least 35 types of candidate DMs are considered, and probably even more given the definitions above. The ground truth annotation was done apparently by one annotator Heeman (1997) but no precise information on the reliability of the annotation is available. The transcribed data has 58,298 words from 34 speakers, and there are 8,278 DMs. Unfortunately, no figure is given for the number of non DM occurrences of the lexical items that can serve as DMs (e.g. sentential *and*), which makes evaluation quite difficult.

The best results of the language model correspond to 533 errors of DM identification, i.e. correct DMs not tagged with a '_D' tag and non DMs that were tagged. With respect to the total number of DMs, this amounts to 6.43% error rate, 97.26% recall and 96.32% precision. While the recall and precision figures concern the identification of DMs as above (see section 1.5.3), accuracy is not computed in the sense of the previous studies (Hirschberg and Litman, 1993; Siegel and McKeown, 1994; Litman, 1996), which we also adopted, that is, the number of correctly identified DMs and non DMs with respect to the total number of candidates. This number is unknown in this study, which does not allow us to compute accuracy or the value of κ. The high values of recall and precision could be related to the high proportion of unambiguous markers in the data, about one word out of seven being considered as a DM. The annotation guidelines, which aren't specified, might also account for the results. According to the authors:

> "Direct comparisons [of the results announced by Hirschberg and Litman (1993) and Litman (1996)] with our error rate of 6.4% are problematic since our corpus is five times as large and we use task-oriented human-human dialogues, which include a lot of turn-initial DMs for co-ordinating mutual belief. In any event, the work of Litman and Hirschberg indicates the usefulness of modeling intermediate phrase boundaries and word accents. Conversely, our approach does not force decisions to be made independently and does not assume intonational annotations as input; rather, we identify DMs as part of the task of searching for the best assignment of DMs along with POS tags, speech repairs and intonational phrases" Heeman and Allen (1999, section 9.3, page 40).

2.4 Conclusion: outline of previous work

It appears from this review—which is also summarized in Table 8.10 at the end of this paper—that the problem of DM identification has been recognized only in a small number of studies. The approaches outlined above make use of a significant range of features, though the limited amount of data did not allow them to apply machine learning to a larger variety of features, for instance lexical features (words surrounding candidate DMs) or speaker-specific information. The ratio between the number of targeted DM types (around 30) and the amount of data (around 1000 examples, or around 10,000 in the latest study) did not allow an in-depth study of the behaviors of DMs, all the more that the studied attempted to identify DMs together *as a class*, which they are probably not. Moreover, the definition of the DM roles and of the

annotation guidelines, as well as inter-annotator agreement data appears to be of secondary importance. The evaluation metrics are not always explicit enough (some studies use only accuracy, others a modified version of it) and the values of baseline scores do not appear clearly, nor is random agreement factored out using *kappa*.

In what follows, we will emphasize the particular behavior of each DM and argue that a larger amount of data and features, as well as the specific treatment of each DM, contribute significantly to increase the accuracy of DM identification.

Chapter 3

Description of the Data

The data used in this study is briefly described in this section, qualitatively and quantitatively. Then, statistical correlations between DMs and speaker characteristics are computed to illustrate for the variability of the data. A comparison with data other studies appears in the next chapter (section 4.3) after the description of the DM annotation process and its results.

3.1 Description of the Corpus and of the Speakers

The ICSI Meeting Corpus of multi-party conversations, which is used here, comprises 75 staff meeting recordings, involving from five to eight speakers (Janin et al., 2003; Morgan et al., 2003). The speech input was recorded individually through separate audio channels and was then manually transcribed. The meetings feature scientific and technical discussions within research groups, dealing with language processing and computer science, and involve native and non-native English speakers. The recordings total about 80 hours, corresponding in transcription to about 800,000 words.

There are 4,519 occurrences of the token *like* and 4,136 of the token *well*, in both sentential and discouse functions, or DM vs. non DM, the separation of which will be discussed in the next chapter (section 4.3). A subsequent annotation effort at ICSI provided segmentation of each channel into about 100,000 individual utterances (temporal and transcript segmentation) which are annotated with dialogue act information (Shriberg et al., 2004). This annotation also provides automatically generated word-level timing—based on 'forced alignment' of transcript with audio—as well as indications of interruptions and unfinished utterances.

The authors of the corpus gathered sociolinguistic information from the speakers such as gender, age, education level, and proficiency in English: native or non-native speaker, region of origin, influences of other languages. Some participants, however, provided incomplete data on the forms. The potential relevance to DM identification of all the available features will be discussed in chapter 6, and empirical evidence based on performance will be brought in chapter 7.

3.2 Characterization of Individual Contributions

This section provides more quantitative data about the speakers' contributions to the corpus, and studies the relatedness of sociolinguistic features with respect to these contributions.

A total of 52 speakers, identified by code names only, contributed to the corpus. The cohort of speakers appears to be well-balanced with respect to the sociolinguistic features. Using the χ^2 test to assess the independence of speaker-related parameters, no significant correlation was

Feature	Value	Number of speakers	Nb. of words	Nb. of utterances	MLU
Gender	female	13	177,761	25,009	7.11
	male	39	615,993	85,528	7.20
Proficiency in English	non native	25	209,630	30,751	6.82
	native	27	584,124	79,786	7.32
Origin	UK	2	11,332	2,079	5.45
	US West	7	98,018	15413	6.36
	other countries	25	209,630	30,751	6.82
	other US	14	256,944	35757	7.19
	US East	4	217,830	26537	8.21
Education	undergraduate	4	14,767	2,604	5.67
	PhD	21	314,416	46,569	6.75
	graduate	21	239,892	34,353	6.98
	professor	6	224,679	27,011	8.32

Table 3.1: Quantitative contributions of speakers to the ICSI Meeting Corpus, relative to sociolinguistic features, sorted by increasing mean length of utterance (MLU). (Note: some speakers provided incomplete sociolinguistic data.

found between gender and education level, gender and age, or gender and origin. However, age and education level are quite expectedly correlated: according to the χ^2 statistic, there are less than 2.7×10^{-5} chances that the present values for age and education are observed under the hypothesis of independence.

The quantity of the individual contributions is however very heterogeneous: the seven most frequent speakers pronounce more than 40,000 words each, and account together for 64% of the data (in terms of number of words, i.e. tokens), whereas the ten least frequent speakers pronounce less than 1,000 words each, and account together for 0.6% of the data. The quantitative contribution to the corpus for each sociolinguistic feature appears in Table 3.1.

When considering the actual contributions of each speaker to the corpus in terms of number of words, nearly all the speaker-related parameters appear to be quite correlated, due to the fact that only a few speakers produced most of the data. For instance, we mentioned that two male professors over 50 years old account for 97% of the words produced by speakers from US East. The χ^2 test shows that age and origin, or gender and origin, are unlikely to be independent. Only gender and age are likely to be independent. Potential correlations must be taken into account when drawing conclusions about the sociolinguistic factors that influence DM use, as explained in Section 8.3.3.

3.3 Frequencies of DMs in the ICSI Meeting Corpus

The frequency of DMs depends a lot on the genre of the data under study, which is made, in the present case, of spontaneous dialogues on a series of research topics. Due to its genre, the ICSI Meeting Corpus appears to be biased favourably towards the presence of DMs *like* and *well*, or at least of the corresponding tokens regardless of their function. It even appears that this data exhibits somewhat higher DM frequencies than predicted by the linguistic literature.

Anticipating somewhat the results of the next chapter (4), which discusses the issue of their disambiguation by human judges, it appears that there are 4,519 occurrences of *like*, of which

2,052 serve as DMs, and 4,136 occurrences of *well*, of which 3,639 serve as DMs[1].

A quick look at the ICSI Meeting Corpus (ca. 800,000 words) shows significant differences in the frequencies of various candidate DMs. The most frequent one is *but* (7,815 occurrences), followed by *like* (4,519 of which 45% are DMs), *well* (4,136 occurrences of which 88% are DMs) and *actually* (1,763 occurrences)[2]. Therefore, the two DMs on which this study focuses have considerable frequency in the corpus, and are also clearly more ambiguous than *but* and *actually*—hence the importance of their disambiguation.

As will be shown below (Table 6.6) the average frequency of DM *like* in the ICSI Meeting Corpus is 0.26%, and the average proportion of DM uses is 45.4%; for *well*, the frequency is 0.46%, and the proportion of DM uses is 88.0%. These values appearing to be quite typical of spoken English, as it appears by comparison to those obtained in a number of descriptive studies.

For instance, Fuller (2003) counted the median rates of six DMs among which *like* and *well*, in two different dialogue contexts, an interview (ca. 24,000 words from six speakers) and an informal conversation (ca. 11,000 words). The median frequencies of *like* are 0.55% in the interview and 0.62% in the conversation setting, and respectively 0.36% and 0.55% for *well*[3].

For another term of comparison for *well*, a number of other figures are provided by a recent study of a corpus of 29 non native English speakers:

> "...there were 788 uses of *well* in the Xhosa English corpus [540,000 words], of which 494 (62.6%) were pragmatic. This is much lower than the 2,199 pragmatic uses of *well* in the New Zealand corpus [420,000 words] (74% of all uses of *well*) (a rate of 0.5 vs. 0.09 per 100 words, which is a considerable difference). Of the uses of *well* in a 50,000 word sample of the London-Lund Corpus, 87.4% (439/502) were pragmatic. Although the Xhosa English corpus yielded a lower frequency than both of these mother-tongue corpora, its use of pragmatic *well* is significantly higher than the 53% of pragmatic usage reported for non-native Spanish speakers of English." (adapted from de Klerk, 2005, pages 1189-1190)

Other DM candidates are moderately frequent or uncommon in the ICSI Meeting Corpus, such as *basically* (457 times), *however* (59), *furthermore* (16), or *moreover* (no occurrence). Regarding *however* and *furthermore*, it is quite remarkable that their frequency is still quite above what is expected from a corpus of spoken dialogues. It is indeed quite unusual to see this DM used in spoken dialogue. In the present data, *furthermore* is used by seven different speakers, both native and non-native. The first and second most frequent speakers never use it, but the third one uses it five times.

As with *furthermore*, *however* is found much more frequently in written than in spoken language. According to Lenk (1998), there are about 50 occurrences of *however* in the spoken language transcriptions from the London-Lund Corpus (500,000 words) and about 225 occurrences in a comparable amount of written texts from the Lancaster-Oslo/Bergen (LOB) corpus. The type of activity should also bias the frequency in the opposite direction: *however* is more frequent in formal settings, such as interviews, as opposed to telephone conversations. And last, the regional variation of English, e.g. American vs. British, has also an influence: according to Lenk's study, "no instance of *however* was found in the spoken American data" (Lenk, 1998, page 150). These biases are not confirmed by the present figures: in the ICSI Meeting

[1] The frequencies of DM use relative to speaker characteristics are shown in Table 6.6 in section 6.3 below.

[2] The exact proportion of DM uses of *but* and *actually* is not known since these tokens were not annotated for this study.

[3] The study also indicates the following observed frequencies for various discourse/pragmatic functions of *well*, in the interview vs. conversation setting: mark insufficiency (57% vs. 51%), mitigate face threat (6% vs. 0%), introduce reported speech or new topic (14% vs. 33%), and delay (23% vs. 14%) (Fuller, 2003, Table 10, page 42)

Corpus, there are 57 occurrences of *however*, of which 34 are produced by native American English speakers—note that one native speaker produced 16 occurrences (the same speaker who produced five of the seven occurrences of *furthermore*) and one non native speaker 14.

It appears therefore that the data used in this study offers an acceptable frequency of occurrence of the two lexical tokens that we focus on, together with enough topic and speaker variety to ensure representativity of training data.

Chapter 4

Disambiguation of DM *Like* by Humans

As a preliminary to the annotation of a substantial amount of data, a study of inter-annotator agreement on the DM identification task was performed, with two main conditions. The first condition made use of the dialogue transcript only, while the second allowed the subjects to listen to the audio recordings if needed. We describe first the experiments and the annotation guidelines (section 4.1), and then the results concerning inter-coder agreement (section 4.2). We also compare these results with those of previous annotation experiments (section 4.3) and provide some technical explanations related to the storage and distribution of the physical annotation files as a language resource (section 4.4)).

4.1 Experiments

Two experiments involving human judges are useful indicators of the reliability of the DM identification task: the accuracy of humans helps assessing the scores obtained later by automatic methods. In our first experiment, judges used only a written transcription of utterances containing *like*. In the second experiment, we explored the improvement of inter-annotator agreement thanks to prosodic information obtained by listening to the meeting recordings.

From a methodological point of view, it is important to define the annotation guidelines in terms that avoid reproducing the criteria that will be used for automatic DM identification, to avoid circularity. Indeed, to take an extreme example, if *well* as a DM were only defined as "occurrences of *well* at the beginning of an utterance", then there would be no doubt that an automatic procedure could reproduce almost exactly the human judgments (provided utterance-segmentation is available) and announce "100% accuracy". The phenomenon to be annotated should therefore be defined in a manner that is as independent as possible from the features planned for automatic detectors, for instance by focussing on the intrinsic definition of the phenomenon or on defining examples as in the experiments below[1]

However, it is likely that from a conceptual point of view the risk of circularity cannot—and should not—be fully avoided. Indeed, there is no possibility to check that automatic annotation method will never make use of the features listed in the annotation guidelines. The goal of automatic annotation being to reproduce as close as possible the reference (human annotation) it is no wonder that the definition of the phenomenon will come under close scrutiny from

[1] A similar concern is raised by (Passonneau and Litman, 1997, section 2.2, page 107): "It is only recently that attempts have been made to quantitatively evaluate how utterance features correlate with *independently justified* segmentations, thereby avoiding circularity" (emphasis added).

programmers which will end up trying to reproduce as close as possible the human criteria as well. The solution is therefore to make sure that the human annotation guidelines circumscribe precisely the phenomenon that is targeted, and that they include non trivial criteria which cannot be immediately reproduced by automatic means.

4.1.1 DM Annotation Based on Written Transcriptions

The first experiment involved six human judges, three men and three women, aged 25 to 40. They were divided into two groups of equal size: one of native English speakers, and one of French speakers with a very good knowledge of English. Each judge annotated a number of utterances containing *like*, which were taken from two different sources: 26 utterances came from the transcription of movie dialogues (US English from *Pretty Woman*) and 49 utterances corresponded to one ICSI meeting. The experiment involved therefore less occurrences of candidate DMs compared to Hirschberg and Litman (1993)—in which two judges annotated 953 occurrences—but comparatively a similar amount of instances of the token *like*, which appeared 61 times in (Hirschberg and Litman, 1993, Table 5, page 517). This is also similar to the number of tokens in the pilot studies of *now* and *well* carried out by Hirschberg and Litman (1993).

The participants were asked to decide for each 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 (certain, reasonably sure, or hesitating) as indicated in the detailed guidelines below. The answers were simply written on paper and then collected by experimenters. Before completing the task, all participants received written indications in English concerning the roles of *like* as a DM, as well as examples of DM and non-DM uses. These instructions, reproduced hereafter, focus on the conceptual definition of the DM role and avoid using low-level features similar to those used by automatic disambiguation methods.

***Like* as a discourse marker**

In English, the word *like* is notoriously ambiguous. It can be a preposition (1), an adjective (2), a conjunction (3), an adverb (4), a noun (5) or a verb (6).

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 a 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.

But *like* can also function as a discourse marker. When it is used as a marker, the primary function of *like* is to make explicit to the hearer that the elements which follow it contain a loose interpretation of the speaker's belief or some uncertainty about this belief. In a lot of cases, the loose interpretation signalled by *like* does not qualify one specific element of the proposition, but applies to a larger compositional unit, such as a verb and its complements (see example 5). Regarding syntax, the discourse marker *like* shows greater flexibility than in other uses: it is not a straightforward preposition, conjunction or adverbial. Here are some examples of like used as a discourse marker:

1. What "Thelma and Louise"? Yeah, it's wicked! Starts of a bit boring. First, *like*, twenty minutes and then it gets good.
2. I know, but it wouldn't be any point if someone wanted to be, *like* a doctor and they got into a nursery place.
3. He goes into ah McDonald's (...) he's *like* can I have breakfast and he's *like*, breakfast and he's *like*, breakfast eleven thirty.
4. Erm, well *like* I usually take the train at about twenty past.
5. Scott said to me if Paul *like* tries to take on Ollie he's just gonna break it up.

Task: the following excerpts contain occurrences of *like* (in **bold**). For every occurrence, write **M** in the margin if you think it is a discourse marker or **O** (other) if you don't. You are also asked to mark your degree of certainty (e.g. **M1**; **O3**).
1 = you are absolutely certain
2 = you are reasonably sure
3 = you hesitate
NB: if there is more than one occurrence of *like* in the same utterance, mark each of them separately.

4.1.2 DM Annotation Using Prosodic Cues

In the second experiment, a group of three annotators (two French speakers and one 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 when needed. This experiment did not include dialogues from a movie, but only from a one-hour ICSI meeting, different from the previous one, and containing 55 occurrences of *like*. Two of the participants had already participated in the first experiment, but the meeting was not the one used in the previous experiment.

The participants received the same set of instructions as in the first experiment, and in addition some explanations 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 about 30 minutes. Access to the recording was provided through a hypertext transcript synchronized to the sound file at the utterance level, a multimedia interface developed for the IM2 project.

The instructions given to the same participants, in addition to the instructions of the first experiment, contained a number of indications from the literature about the prosodic behavior of *like* as a DM:

Disambiguation of *like* using prosodic cues

When *like* is used as a discourse marker, it has several specific prosodic specificities. Notably, *like* as a discourse marker is unstressed and usually phonologically reduced. Moreover, if *like* is followed by a brief pause, the phonological separation from the adjacent discourse unit suggests that the form is not syntactically integrated within this unit; hence a discourse marker interpretation becomes plausible.

In this second test, you will be able to pay attention to prosodic cues as well. Feel free to listen to the tape recording as many times as you need.

Task: the following excerpts contain occurrences of *like* (in **bold**). For every occurrence, write **M** in the margin if you think it is a discourse marker or **O** (other) if you don't. You are also asked to mark your degree of certainty (e.g. **M1**; **O3**).
1 = you are absolutely certain
2 = you are reasonably sure
3 = you hesitate
NB: if there is more than one occurrence of *like* in the same utterance, mark each of them separately.

4.2 Results and discussion

In the first experiment, based on transcripts only, the level of inter-annotator agreement was quite low for the natural dialogues of the ICSI corpus ($\kappa = 0.40$), and barely acceptable for the movie ($\kappa = 0.65$) see Section 1.5 for the intepretation of the *kappa* (κ) score). In the second experiment, with the help of prosodic cues, inter-annotator agreement increased, and the annotation became much more reliable, at $\kappa = 0.74$. Therefore, the identification of DMs when the audio recordings are available appears to be an empirically valid task. The necessity of prosodic information appears quite clearly, and confirms previous observations on the human annotation of other discourse phenomena, in particular discourse segmentation (Grosz and Hirschberg, 1992; Hirschberg and Nakatani, 1996).

These results shed an interesting empirical light on a number of predictions that we can make independently of the the experiments. First, it appears that DMs are easier to annotate in pre-planned dialogues, because such dialogues are less ambiguous than the natural ones: the agreement level reached for the movie transcript is much higher than for the ICSI meeting (0.65 vs. 0.42). So, even if movie dialogues are meant to reproduce the naturalness of naturally

occurring dialogues, they are much less ambiguous, mainly because they reflect the global communicative intention of only one person (the author).

The second hypothesis we tested concerned the difference between the abilities of native and non-native speakers 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 $\kappa = 0.40$ vs. $\kappa = 0.43$ for the meeting corpus. We conclude that non-native English speakers with a very good command of English are just as reliable as native English speakers to annotate DMs.

We also tested the correlation between the degree of certainty of annotators (captured by their use of numbers 1,2 and 3 to indicate certainty) and their level of agreement. However, we did not find any significant correlation. The capacity of human judges to evaluate their own intuition does not seem to be very high for this task. The subjects were more confident when they were able to use prosodic cues: the percentage of answers given with maximal certainty grew from 45% to 60% and from 65% to 87% for two annotators.

When looking more closely at the utterances upon which annotators do not agree, 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 as DMs all occurrences of the type: *sounds like*, *seems like*, *feels like*. This observation is not so surprising if we bear in mind that the pragmatic uses of *like* seem to have emerged, historically, through a grammaticalization process (see also section 1.2 above). Andersen hypothesizes that "the pragmatic marker *like* originates in a lexical item, that is, a preposition with the inherent meaning *similar to*" 2001, page 294. Thus, a small number of DM occurrences remain ambiguous, since they are at the boundary between DM and non-DM use, both interpretations being equally possible. These cases, about 0.5% of the total, are not used in the following experiments, much like Hirschberg and Litman (1993) who also removed 'non classifiable' candidate DMs from their data sets.

These human annotation experiments quantified the level of inter-annotator agreement and confirmed the necessity of prosodic cues for reliable detection of DM *like*. The annotation task can be accomplished at a reasonable performance level even by untrained annotators. The inter-annotator agreement scores offer an initial upper-bound for automatic performance level, which should not be expected to reach much higher levels.

In the following experiments with automatic annotation, the full set of data was annotated by the two authors for about 50% of the data, and by one author for the remaining 50%. Disagreements were removed from the training/test data, and in addition annotators were allowed to mark instances as 'uncertain', which also led to their removal from the data.

4.3 Comparison with Other Annotated Corpora

We summarize here, for comparison purposes, the corpora and the methods for ground truth DM annotation used in other studies of DM identification.

4.3.1 Ron Brachman's Talk

Hirschberg and Litman (1993) describe first a study based on a radio show transcript containing 100 occurrences of *now*, for which "the authors determined separately, and by ear, whether individual tokens were discourse or sentential usages" (page 511)—discourse uses are defined as signalling directly the structure of discourse; 37 instances are tagged as sentential, and 63

as discourse. Conclusions derived from this analysis were tested on a corpus containing 52 instances of *well*, with 27 discourse uses.

However, the central part of their study is based on a corpus of single-speaker transcribed speech from a keynote address at an AI conference in 1986 (75 minutes speech with ca. 12,500 words in transcription). Their data contains 953 occurrences of cue phrase candidates, of 34 types. Their distribution is quite uneven: the most frequent token, *and*, occurs 320 times (33.6%), while 18 tokens occur less than 10 times each, and a number of well-discused tokens do not occur at all (such as *anyway* or *yet*).

The authors indicate that tokens appear more frequently in the introductory remarks than in the remainder of the talk, which shows that they are more characteristic of spoken interaction that of formal spoken monologues. The tokens were classified as 'discourse' or 'sentential' or 'ambiguous' by two annotators (the two authors). There were 59 ambiguous tokens (i.e. 6.2% of the total) and 16 tokens on which the annotators disagreed, partially or completely (i.e. 1.7%). The numbers of disagreements and ambiguous tokens were about four times higher on the tokens used as conjunctions (*and*, *or*, *but* than on the other tokens. Among the classified tokens, 341 (38.8%) had a discourse (DM) role and 537 (61.2%) had a sentential (non DM) role (Hirschberg and Litman, 1993, page 518).

The same data containing 953 DM candidates was used by Litman (1994, 1996), who also considered several subsets: the subset of 878 classifiable tokens, i.e. excluding those marked as ambiguous, and the subset of 495 classifiable non-conjuncts, i.e. excluding also the potential conjunctions *and*, *or* and *but*.

Siegel and McKeown (1994) use roughly the same corpus as Hirschberg and Litman (1993), more specifically 1,027 training examples from "a transcript of a single speaker speech, preceeded by introductory remarks by other speakers [NB: this does not appear in (Hirschberg and Litman, 1993, page 517)], in which each occurrence of the words in Table 1 [NB: same list as Hirschberg and Litman (1993)] has been manually marked as to its meaning by a linguist. [Note 6:] We used one linguist's markings, whereas Hirschberg and Litman (1993) used and correlated the judgments of that and another linguist, discarding those cases in which there was disagreement. Thus, the data we used was slightly more noisy than that used by Hirschberg and Litman (1993)." They also use a slightly larger portion of the transcript than Hirschberg and Litman (1993) which explains the slightly larger set of instances. Therefore, they have 1,027 classifiable (unambiguous) occurrences, of which 407 are DMs (39.6%) and 620 (60.4%) are not DMs.

4.3.2 The Switchboard Corpus

Another attempt to annotate the pragmatic occurrences of DMs comes from research on the Switchboard corpus of telephone conversations. A set of DMs—among which *well*, *you know*, *like*, *so*, and *actually*—was hand-annotated, along with "fillers", "explicit editing terms", and coordinate conjunctions (Meteer, 1995). The instructions given to human annotators were somewhat underspecified, maybe lacking enough theoretical grounding for the definition of DMs (a fact that was acknowledged by the authors themselves). For instance, DMs were simply considered to have "more semantic content than fillers (although not always much)". The annotation difficulties are especially visible for *like* and *so*. When in doubt, annotators were told to use the following strategy: "if the speaker is a heavy discourse *like* user, count ambiguous cases as DMs, if not, assume they are not." Finally, the annotation of some DMs was not completed: the authors noted that *actually* "proved impossible for the annotators to mark consistently and was jettisoned as a DM part of the way through" (Meteer, 1995). The data thus produced has not been used (to our knowledge) to develop an automatic DM identifier. Our previous experience with this data confirmed that the DM annotation available

on Switchboard is not reliable enough to be used as ground truth data (Zufferey and Popescu-Belis, 2004, Section 6).

4.3.3 The Trains Corpus

Heeman (1997); Heeman et al. (1998); Heeman and Allen (1999) use a substantially different corpus, namely transcripts of the TRAINS corpus of human-human planning dialogues. The corpus comprises nearly 100 dialogues with almost 60,000 words, among which there are 8,278 DMs. The data was POS-tagged by human annotators, and DMs were marked in this process as well. The full list of DMs is not available, though examples of each of the four possible classes of DMs are given: single word acknowledgments (tagged with AC), discourse conjuncts (CC_D), discourse adverbials (RB_D), interjections (UH_D). It seems that verbs used as DMs, or multi-word DMs are not considered in this study. The list of DM examples illustrating the four discursive tags contains 23 items (Heeman et al., 1998, Table 2), but it is not clear whether this is the full list or not (a shorter list is quoted by Heeman and Allen (1999, page 9)). Unfortunately, no figures are given regarding the number of annotators and their agreement.

4.4 Technical Note on the Resulting Corpus and its XML-based Annotation

Technical issues related to annotation formats and tools must be considered by programmers and by those who want to reuse the present DM annotation, which is available freely at `www.issco.unige.ch/projects/im2/mdm/data/discourse-markers`. The resources consists of 8,655 occurrences of *like* and *well* of the ICSI Meeting Corpus.

The initial format used for processing the transcript is the dialogue act annotation format generated by Shriberg et al. 2004: each line contains the transcript of one utterance, plus the indication of the speaker, start time and end time of each word, the dialogue act associated to the utterance, etc. We first created a copy of each utterance transcription for each occurrence of *like* and *well* that it contains and generated a table for human annotators, with an empty column that must be marked '1' for a DM use and '0' for a non-DM use (the number '2' can be used to signal a token that is difficult to classify). The position within the utterance of the token is indicated in a previous column, e.g. '2' means "second occurrence of *like* in the utterance".

An XML export format was defined in order to re-use previous annotations, along with software that enables import, export, and merging of annotations. Human annotation of the resource is first imported from the tabular format, then merged with previous annotations stored as XML, while differences, if any, are reported. The merge operation is essential since it allows incremental annotation and correction of the data.

The full annotation can be re-exported as XML, or processed by the feature extractors described in Section 6, in preparation for automatic annotation. This generates the final working format, that is, a list of instances for C4.5 training, containing the values of the features plus the resulting class, DM or not DM. The import→merge→export cycle can be repeated each time the annotation is revised or enriched. This series of operations, including processors (Perl scripts) and resulting files, is summarized in Figure 4.1.

We defined a simple XML stand-off annotation format for the export data. The DM annotation is separated from the transcripts, which allows its distribution without disclosing the ICSI Meeting Corpus. The annotation file is divided into `<dialog>` elements corresponding to each meeting, composed of `<dmtoken>` elements that provide the DM/non-DM information for each token. For example, the beginning of the DM annotation file is:

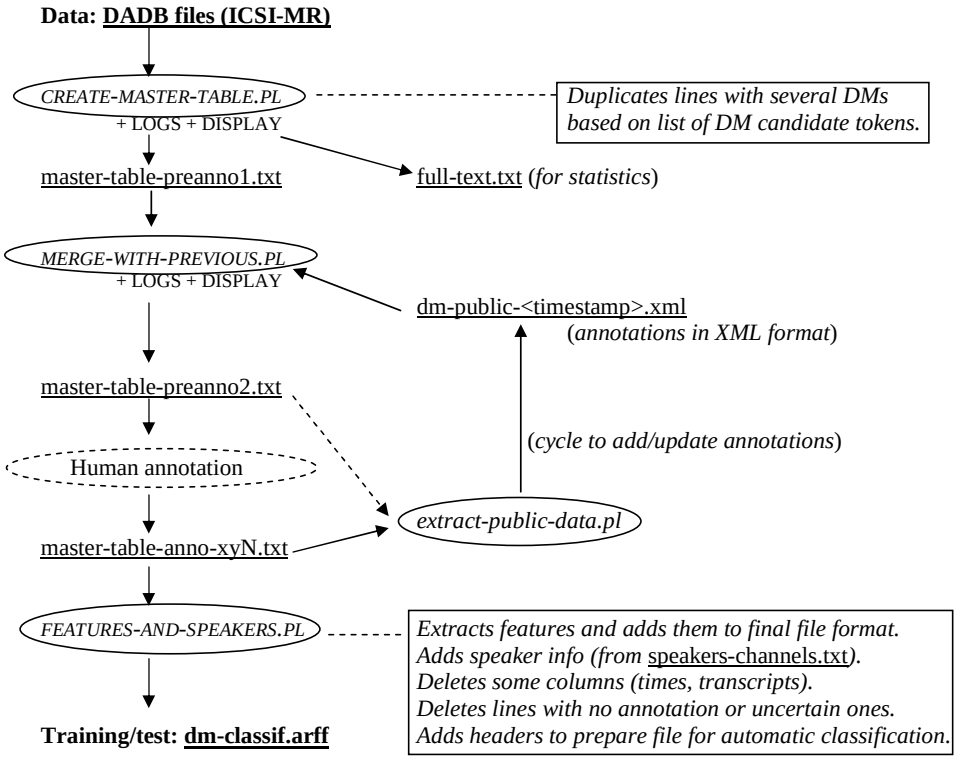

Figure 4.1: Processing DM annotations: processors and resulting files.

```
<discourse_markers>
  <dialog  id='Bdb001'>
    <dmtoken  type='like'  position='1'  channel='4'
              startTime='20.56'  endTime='21.89'  dm='0'/>
    ...
```

This encodes the fact that in dialogue 'Bdb001' (`id` of parent `dialog` element), the first occurrence (`position='1'`) of *like* (`type='like'`) in the utterance starting at 20.56 seconds and ending at 21.89 seconds on channel 4 is not a DM (`dm='0'`). We do not use the timing of the token itself, since this is not fully reliable, as it was computed automatically. The DM annotation we provide (at `www.issco.unige.ch/projects/im2/mdm/data/discourse-markers`) contains all the necessary information to recover the DM data, if researchers have access to the ICSI Meeting Corpus transcripts.

Chapter 5

Disambiguation of DM *Like* by a POS 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.

5.1 Previous Attempts

The use of a POS tagger was in fact adopted in at least two of the previous studies of DMs. Litman (1996) used the POS tags assigned to the candidate DMs by Church's tagger as a feature for DM identification. It appears however that the contribution of this feature to automatic identification is significantly smaller than that of preceding orthography (37.7% error rate for 'POS' vs. 18.8% for 'preceding' in (Litman, 1996, Table 6, page 75)), and its addition to all the other features does not increase performance (18.8% for 'preceding' vs. 18.7% for 'text' in the same table). It would have been interesting to test, in this study, the accuracy of the POS tags assigned automatically, as errors in POS tagging may render such a feature inoperant.

Heeman and Allen (1999) incorporated POS tagging, DM identification and utterance segmentation into a unique probabilistic language model for automatic speech recognition (see section 2.3.4 above). DM identification becomes thus a POS tagging problem. One of the most interesting results is that the number of POS errors of the language model decreases when DMs are included in the model by 2.5% (Heeman and Allen, 1999, Table 7, page 18)—note however that "to ensure a fair comparison, [this study did] not penalize POS errors that result[ed] from a confusion between discourse and sentential usages." Due to the order of integration into the language model, no scores are available for DM identification without POS tagging. When intonational phrase detection is added to POS tagging, DM identification improves (by ca. 7%), but when speech repair detection and correction is further added, DM identification is degraded (by ca. 10%) (Heeman and Allen, 1999, Table 10, page 35).

These results thus emphasize the dependency between POS tagging and DM identification, which appear to be mutually related tasks, with no obvious order of priority.

5.2 DM Tagging Using QTag

We experimented on about a quarter of the ICSI Meeting Corpus using QTag, a general-domain probabilistic POS tagger for English (Mason, 2000). QTag uses a variant of the Brown/UPenn

tagsets, and was trained on a million-word subset of the BNC (written material) and is available at: `http://web.bham.ac.uk/o.mason/software/tagger/`. The tagger assigns one of the following tags to occurrences of *like*: preposition ('IN', 1,412 occurrences), verb ('VB', 509), subordinate conjunction ('CS', 134), general adjective ('JJ', 52), and general adverb ('RB', 9).

These tags should further be interpreted in terms of DM vs. non-DM uses. A simple attempt to do so is to use the tagger as a filter to remove verbal occurrences of *like*. Hence, a 'VB' tag could be interpreted as non-DM, and all the other tags as (possible) DMs. Unfortunately, evaluation shows that such a filter is totally unreliable: its recall score is 0.77, precision is 0.38, accuracy 44%, and kappa only 0.02, i.e. a nearly random correlation. No other interpretation of the POS tags leads to better overall results. The highest scores are obtained when selecting only adjectival uses of *like* (tagged 'JJ') as potential DMs: recall is of course very low, but precision is 0.74, which means that the 'JJ' tag could be used as a clue for the presence of a DM use of *like*.

5.3 Discussion

The main reason for the failure of the tagger to detect DM uses of *like* in the ICSI Meeting Corpus is probably the fact that it was not trained on speech transcripts, where DM *like* is quite frequent. A tagger trained on spoken dialogues could use for instance the interruptions or pauses that sometimes appear around DM uses of *like*, which would prevent it from marking some of those occurrences as 'VB'. As mentioned, Heeman et al. (1998) showed that when specific tags are assigned to DMs, and tagging is done in the process of speech recognition, both the quality of tagging and the correct identification of DMs are significantly improved.

The idea of training a POS tagger on the ICSI Meeting Corpus data by adding a 'DM' label to the tagset does not seem as efficient as the decision tree learning described below, since the most widespread tagging methods do not learn transparent rules, as we will do. An exception, Brill's tagger, makes use of collocation rules only, which we exploit too, but rule-based POS tagging is generally forward-looking (from $word_n$ to $word_{n+1}$), and several passes are needed to propagate information backwards—-while we use below both $word_{n-1}$ and $word_{n+1}$ collocations thanks to a decision tree using the corresponding features.

Chapter 6

Analysis of Features for DM Identification

The proposed automatic method for disambiguating DM candidates relies on a variety of features which extend those used by Litman (1996) and which are used by automatically trained decision tree classifiers, thanks to a large set of training examples. In this chapter, a set of relevant surface features that can be extracted with lightweight processing is outlined.

The method focuses on surface features only, since deeper analyses of an utterance seem to require in most cases the prior identification of DMs. For instance, it would not be realistic to assume the availability of a parse tree or of a deep semantic analysis of an utterance, as they require knowledge of the DM function; conversely, if they were available, information about the role of candidate DMs could be derived from them quite easily.

The following sets of features are described: collocation-based criteria (6.1), prosodic and positional criteria (6.2), and speaker-related features (6.3). Although these features generalize to a larger set of DMs, the focus here is their application to *like* and *well*. The 'token' feature is also used, which represents the nature of the candidate token, here *like* or *well*, as used by Litman (1996). The use of the 'token' feature for training allows the two types of DMs to be processed differently. Conversely, when this feature is not made available, a single classifier is built, thus forcing generalization. The relevance 'token' feature will thus show the relevance (or not) of processing candidate DMs differently according to their nature.

6.1 Lexical Collocations

In many cases the word immediately preceding or following a DM candidate influences the likelihood that the token is a DM or not. These features are quite naturally DM-specific, as each DM takes part in specific lexical constructions. A first idea is to build 'collocation filters' out of these rules (Zufferey and Popescu-Belis, 2004), considering that some collocations tend to occur with DM uses while others with non DM uses. However, a more robust solution is based on automatic learning of such filters from the data itself, which removes the need for manual definition of the filters and ensures they maximize DM identification accuracy.

This section highlights the main collocations that are found through a simple statistical analysis of the data. These will be compared later on to automatically derived collocations. Collocations can be distinguished according to their position: the candidate DM can be the first or the second term of the collocation. Collocations with more than two words are observed as well (e.g. *well you know*).

6.1.1 Examples

Many studies have observed that *well* as a DM appears frequently in collocations, for instance:

> Of the 494 pragmatic uses of *well* in the XE corpus, 35% ($n = 175$) were immediately followed by a personal pronoun (105 *well I'm/I've*; 34 *well you*; 13 *well he*; 11 *well they*; 12 *well we*). In comparison, the NZ corpus had 33% ($n = 733$) collocations of *well* directly with a personal pronoun. This therefore suggests a similar strongly interpersonal role for well in both NZE and XE. Other notable collocates in XE included signs of agreement (29 *okay/oh/yes/yeah/ja*), and common formulaic expressions such as *well I think* (12), *well you know* (11) and *well you see* (5). This was considerably lower than the distribution of such expressions in the NZ corpus (169 *okay/oh well*, 45 *yes/yeah* and 79 *well I think/you know/you see*) (de Klerk, 2005, page 1190).

More generally, when *well* is used as a DM to mark a change of topic, it is nearly always used in a cluster of markers such as *well now* or *oh well*: so, *now* and *oh* appear to be indicators of DM uses of *well*, appearing as first terms of collocations. Similarly, when used as a DM to close a topic, *well* can very often be found in clusters such as *OK well* or *well anyway*, therefore *anyway* and *OK* could be indicators of DM uses of *well*.

In contrast with collocations that tend to occur with DM uses, some other collocations tend to occur with non DM uses. For instance, this is the case when *like* is used in constructions such as *I like*, *seems like*, *feels like*, *just like*. Similarly, *well* is not a DM in constructions such as *very well*, *as well*, *quite well*.

Two extensions of collocations could also be explored. First, we considered non adjacent words as features, but linguistic analysis provided little evidence for any such long-distance correlations with our DMs, unlike the case of discourse connectives (Hutchinson, 2004a). Second, we also tried two-word collocations—i.e., tri-grams including DM-candidates—but further analysis and the comparison of Table 6.1 with similar ones for tri-grams, showed that they did not bring additional information.

6.1.2 Observations on Data

To study the efficiency of the collocation based features, one can examine how they are used by an automatically trained classifier that is allowed to choose the most relevant features for DM identification, as in Section 8 below. However, it is also possible to conduct an *a priori* analysis of the discriminative power of collocations, as shown in Table 6.1. Only collocation features with high discriminative power are listed, that is, those with a ratio far from 1, either very small (say smaller than 1/9) or very big (say bigger than 9/1)—that is, collocations that appear either predominantly with DMs, or predominantly with non-DMs. In addition, only collocations that are frequent enough, and that have a reasonable linguistic interpretation, are kept in the list.

Other measures of *a priori* discriminative power appeared to be less suitable. Mutual information between two words (Manning and Schütze, 1999, chap.5) does not seem relevant here for several reasons. First, one of the words in the collocation is always fixed (the DM-candidate token); second, collocations that appear infrequently have maximal mutual information, but little generality; and third, the available data is still too sparse to draw reliable conclusions from mutual information values. Another score for the relevance of each collocation could use the four numbers of joint occurrences of the token (t) and an indicator word (i): $(t \wedge i)$, $(\neg t \wedge i)$, $(t \wedge \neg i)$, and $(\neg t \wedge \neg i)$. From these values, probabilities of the respective classes can be computed, and

Collocation	Nb. of occ.	% DM	% non DM	Indication
like that	579	9	91	non DM
like a	348	41	49	none
like the	293	42	58	none
like to	232	4	96	non DM
like this	192	13	88	non DM
like you	156	59	41	none
like uh	98	47	53	none
like I	95	64	36	none
like it	86	22	78	none
like it's	66	30	70	none
like if	62	76	24	none
something like	464	8	92	non DM
it's like	267	66	34	none
things like	202	7	93	non DM
seems like	185	4	96	non DM
of like	173	79	21	DM
would like	136	3	97	non DM
is like	126	70	30	none
was like	100	88	12	none*
I like	92	10	90	non DM
looks like	90	3	97	non DM
sounds like	80	13	88	non DM
have like	76	95	5	DM
just like	74	57	43	none
look like	60	2	98	non DM
stuff like	59	7	93	non DM
I'd like	56	2	98	non DM
know like	50	80	20	none
well I	445	96	4	DM
well the	167	95	5	DM
well we	155	97	3	DM
well you	156	94	6	DM
well it's	135	98	2	DM
well that's	125	99	1	DM
well it	117	75	32	none
well if	95	89	11	none
well yeah	72	94	6	DM
well but	66	97	3	DM
as well	204	7	93	non DM
very well	61	3	97	non DM
uh well	58	90	10	DM
oh well	44	100	0	DM
say well	45	91	9	DM
pretty well	39	3	97	non DM

Table 6.1: Most frequent collocations for *like* and *well* observed in the ICSI Meeting Corpus and their hypothesized role as DM indicators, depending on their meaning and the DM to non-DM ratio (no indication is hypothesized for *was like* despite the observed ratio, given the manifest ambiguity of this collocation.

finally a value of κ that expresses the correlation between the token and the indicator. However, this score would be less interpretable than the ratio we used.

A possible role of the collocation features is to use only the most reliable 'exclusive' collocations, in other words, keep recall close to 100%, and try to maximize precision (Zufferey and Popescu-Belis, 2004). This resembles Hutchinson's 2004b data collection method from the Web (see Section 2.2.4). This method could help reducing the number of tokens submitted to human annotators, by eliminating those that are certainly not DMs. The filters presented in Table 6.1 reach 80% precision for *like*, with nearly 100% recall. This would already be enough to reduce the burden of human annotators, though we did not use them for the moment in the annotation effort.

6.1.3 Representation of Lexical Features for DM Identification

Given the hypothesized roles of lexical collocations for DM identification, this type of information must be available for training the DM classifiers. As a result of training, the DM classifier will use the lexical features that appeared to be the most informative for DM identification, depending of course on the training method.

Two main parameters influencing the nature of lexical features will be tested empirically. The first one is the number of words preceding and following the candidate DM that will be considered as features for classification, or in other terms the size of the window surrounding the candidate DM[1]. If the N preceding and following words are considered for disambiguation, the size of the window is $2N$.

The second parameter is the minimal occurrence frequency required for words to be used as features to train DM classifiers. A trade-off must be found between the use of all (or most) of words co-occurring with candidate DMs, which might lead to over-specialization of the classifier with respect to the training data, and the use of only the most frequent words, which may be to general and fail to classify many candidate DMs. In other terms, only words appearing more than F times in window of size $2N$ around the candidate DM will be used for classification. For $N = 1$ (i.e. using $word_{-1}$, $word_{+1}$), the thresholds of $F = 3$, $F = 10$ and $F = 20$ correspond respectively to 360, 150 and 90 words as possible values of the features features. For $N = 2$ (i.e. the [-2; +2] lexical window) these values are respectively 700, 250 and 160.

From a formal point of view, there are two logically equivalent possibilities to encode the lexical features:

1. Use *one variable for each word position* with respect to the DM candidate, the possible values of the variables being the possible words observed surrounding the DM candidates (above a certain frequency threshold)[2]. In other words, this option uses $2N$ variables $word_{-N}$, $word_{-N+1}$, ..., $word_{-1}$, $word_{+1}$, ..., $word_{N-1}$, $word_N$ each of which having several hundreds of possible values. The classifiers are expected to construct rules such as, for *like*, (($word_{+1}$='to') $\Rightarrow$ (non DM)) or for *well*, (($word_{-1}$='oh') $\Rightarrow$ DM) (cf. Table 6.1 for likely collocations). In case only the first preceding and following words are used, as in most of the experiments below, only $word_{-1}$ and $word_{+1}$ will be used. The features can thus be summarized as:

   $word_{-N}$, ..., $word_{-1}$, $word_{+1}$, ..., $word_N$. Possible values: for each feature, the list of words occuring more than F times in a window of size $2N$ around candidate DMs in

[1] For instance, Siegel and McKeown (1994) indicated that only the words immediately following the candidate DM were useful, an observation echoed by de Klerk (2005, page 1190): "DMs typically act as a guide to addressees as to how to react to what is about to be said, rather than acting retrospectively on what has already been said."

[2] Technically speaking, the list of possible values could be different for each variable based on observed words in the respective position, but we believe this will decrease the generality of classifiers built using these features.

the whole ICSI Meeting Corpus; if there is no such word in the utterance (e.g. for $word_{-1}$ in case the candidate DM is utterance initial) then the value of the feature is 'absent'; if it is a word that is not in the list of most frequent ones, then the value is 'other'. Hypothesized role: following Table 6.1, some lexical items are reliable indicators of DMs, while others indicate non DMs, and other ones are neutral (the classifier is expected to find some of the conclusions expressed in the last column of Table 6.1).

2. Use *one variable for each possible word* (from the list of the most frequent ones), the value of which indicates the position of this word with respect to the candidate DM: +1 if it is the following word, -1 if it is the preceding word, etc. The null value should be used if the word is not present in a window of size $2N$ around the DM candidate—so for each candidate DM, most variables (except at most $2N$ ones) will have the null value. However, this convention poses some problems to the "linearity" of these features.
   This option increases the number of variables, but greatly reduces the number of their possible values. The only restriction with the respect to the previous solution is that if a given word is present both before and after the candidate DM, only one value can be encoded in this system, so a system of priorities should be established, for instance coding first the instances tht are closer to the DM candidate, starting with the preceding words (shown to have more influence). It is also quite unlikely, though not impossible) that a candidate DM is surrounded twice by the same world. The classifiers are expected to construct rules such as, for *like*, (position('to')=+1 ⇒ (non DM)) or for *well*, (position('oh')=-1 ⇒ DM) (cf. Table 6.1 for likely collocations. The features can be summarized in this case by:

   position($word_i$): where $word_i$ is one of the words appearing more than$F$ times in a window of size $2N$ around the candidate DM. Possible values: for each feature, the relative position with respect to the candidate, or 0 if the word is absent around that particular candidate. Hypothesized role: the same as for the above encoding.

Although logically equivalent, these two methods are not treated identically by machine learning algorithms, due to internal specificities. For instance, when learning decision tree classifiers, the first solution tends to generate fewer nodes but with many more branches, while the second can only generate nodes with at most three branches (-1, 0 or 1), so decision trees might have much more nodes (or, if they have to be pruned, they will remain compact).

6.2 Position and Prosody

Insights from linguistic and computational studies show that position and prosody indicate whether a token functions as a DM or not. For instance, *well* as a DM appears nearly always at the beginning of an utterance or of a prosodic unit. In other cases, it is a non-initial position that marks a DM, as observed by Aijmeer 2002, page 30: "some of the discourse particles [...] (*actually*, *sort of*) can, for instance, be inserted parenthetically or finally, often with little difference in meaning, after a sentence, clause, turn, tone unit as a post-end field constituent." As for prosody, according to Schiffrin 1987, page 328, "[a discourse particle] has to have a range of prosodic contours, e.g. tonic stress, and [must be] followed by a pause [or a] phonological reduction". Not all the prosodic features are however easy to extract automatically as features: for instance, Hirschberg and Litman (1993) use a manual annotation of prosodic accent[3].

[3] At the utterance level, however, Shriberg et al. (1998) were able to automatically annotate "duration, pause, F0, energy, [and] speaking rate".

Token	Role	Nb. occ.	Utt. initial	Utt. final (compl.)	Utt. final (interr.)	Utt. final (total)
Like	DM	2,052	342	9	138	147
	non DM	2,467	92	69	58	127
	both	4,519	434	78	196	274
Well	DM	3,639	2,755	117	249	366
	non DM	497	10	208	2	210
	both	4,136	2,765	325	251	576

Table 6.2: Quantitative data related to positional features.

As mentioned above, the position of a candidate DM in the utterance is another important factor in DM identification (Hirschberg and Litman, 1993; Litman, 1996; Heeman and Allen, 1999). A descriptive study on the use of DMs in two different conversational settings found that *like* appeared mainly in turn or utterance medial positions, and *well* appeared mainly in turn initial position, possibly following another DM, but never in utterance medial position (Fuller, 2003, Table 9, page 41)[4].

6.2.1 Observations on Data

In the present corpus, the statistics indicated in Tables 6.2, 6.3, 6.4 and 6.5 were observed regarding prosody and position. These figures where obtained through forced-alignment of the audio recordings with their manual transcripts, using the automated method described by the creators of the corpus and of its transcriptions (Janin et al., 2003; Morgan et al., 2003; Shriberg et al., 2004). Therefore, these figures are only approximations of the real timing of the words and of the pauses between them—in most cases there is no pause between the words of a continuous prosodic phrase—which are rounded to the nearest ten milisecond value. Although the timing in seconds is provided with three decimals, the last one is always the same across an utterance, therefore the timing of the words is rounded to the second decimal (tens of miliseconds).

The observed values do not show very strong correlations between features and the DM character. It appears for instance from Table 6.2 that very few DM *like* are utterance finals in a completed utterance, and that proportionally, there almost four times more utterance initial *like* that are DMs than non DMs. Similarly, *well* as a non DM is extremely rare in utterance initial position or at the end of interrupted utterances.

As for prosodic/temporal information, Tables 6.3, 6.4 and 6.5 indicate that there are a number of differences regarding DMs vs. non DMs in terms of the duration of the pauses surrounding them and their duration. Most of these parameters appear to differ between DMs and non DMs: the most notable differences appear for the pause before the token (for both lexical items) and the pause after the token (for *like* only). The differences with respect to duration are less marked, and there is almost no differences regarding the pause after *well*.

In fact, it is not always easy to assess the importance of these differences: for instance, the pause before non utterance-initial DM *like* is on average 59 ms vs. only 12 ms for non DMs,

[4]The complete results for the informational conversation setting vs. interview setting show that of the 83 occurrences of DM *like* in conversations (resp. 198 in interviews), 60% are turn medial (resp. 48%) and 38% are utterance medial (resp. 46%); a negligible amount were turn initial, second after a DM, alome or turn final. As for *well* in the two settings (67 and 73 occurrences), 69% were turn initial or second initial following another DM in the conversation setting (resp. 55% for the interview), and 28% were turn medial (resp. 43%). A negligible amount of *well* were utterance medial, alone or final. Note also that de Klerk (2005, page 1190) indicates that "of all occurrences of *well*, 32.4% (158) were turn initial (in a further 48 cases, *ja*, *um* and *okay* preceded *well*)."

Token	Role	Nb. occ.	Average duration of pauses **before** non initials (ms)	Confidence interval (ms)	Standard deviation (ms)	Nb. of non initials	Nb. of occ. with pause >0
Like	DM	2,052	59	9	190	1,710	287
	non DM	2,467	12	3	77	2,375	136
	both	4,519	31	4	138	4,085	423
Well	DM	3,639	306	74	1,119	884	343
	non DM	497	23	11	129	487	38
	both	4,136	206	48	912	1,371	381

Table 6.3: Prosodic/temporal features: pauses before candidate DMs.

Token	Role	Nb. occ.	Average duration of pauses **after** non initials (ms)	Confidence interval (ms)	Standard deviation (ms)	Nb. of non initials	Nb. of occ. with pause >0
Like	DM	2,052	114	11	251	1,905	594
	non DM	2,467	42	6	158	2,340	321
	both	4,519	74	6	208	4,085	915
Well	DM	3,639	76	8	234	3,273	651
	non DM	497	64	26	223	287	48
	both	4,136	75	8	233	3,560	699

Table 6.4: Prosodic/temporal features: pauses after candidate DMs.

but the standard deviation for these values is very high: 190 ms and 77 ms respectively (see Table 6.3). This shows that the duration of the pause before *like* varies considerably around the respective average values, though this duration is *on average* larger for DM than for non DMs (this is an aspect of the marked prosodic contour of DM *like*). The large variation of the values around the average (the large standard deviation) suggests that the pause-before may help to identify DMs from non DMs only in a very limited way, because there is no value that would provide a clear-cut separation between DMs and non DMs (for instance the center of the [12 ms; 59 ms] interval).

It is also important to note that, when considering the confidence intervals for these values, the low values that are observed do not indicate that most of the observation are really so close to the average value. For instance, the average duration of the pause before non initial, non DM *like* is 12 ms with a 95% confidence interval of ±3 ms (Table 6.3). This of course does not mean that 95% of the tokens are preceded by a pause in the [9 ms; 15 ms] interval! Statistically, the meaning of this confidence interval is that if 100 samples of 2,375 non DMs *like* were drawn from similar dialogues, 95 of them (on average) would exhibit an average pause before in the [9 ms; 15 ms] interval, which is quite a different fact. Moreover, as shown in the last column of Table 6.3, only 136 of the observed 2,375 tokens are preceded by a nonzero pause, a fact that contributes to reduce the size of the confidence interval, but increases standard deviation. The fact that the spread of the values around the average is better characterized by the standard deviation than the confidence intervals appears also in the analyses of the duration of the tokens discussed below (see Figures 6.1 and 6.2).

Similarly, there is a marked difference in the average pause before non utterance-initial DMs *well*, which is 306 ms, and the pause before non DMs *well*, which is only 23 ms. The standard

Token	Role	Nb. occ.	Average duration (ms)	Confidence interval (ms)	Standard deviation (ms)
Like	DM	2,052	225	3	75
	non DM	2,467	211	3	75
	both	4,519	217	2	75
Well	DM	3,639	200	3	95
	non DM	497	259	9	100
	both	4,136	207	3	97

Table 6.5: Quantitative observations of prosodic/temporal features: durations.

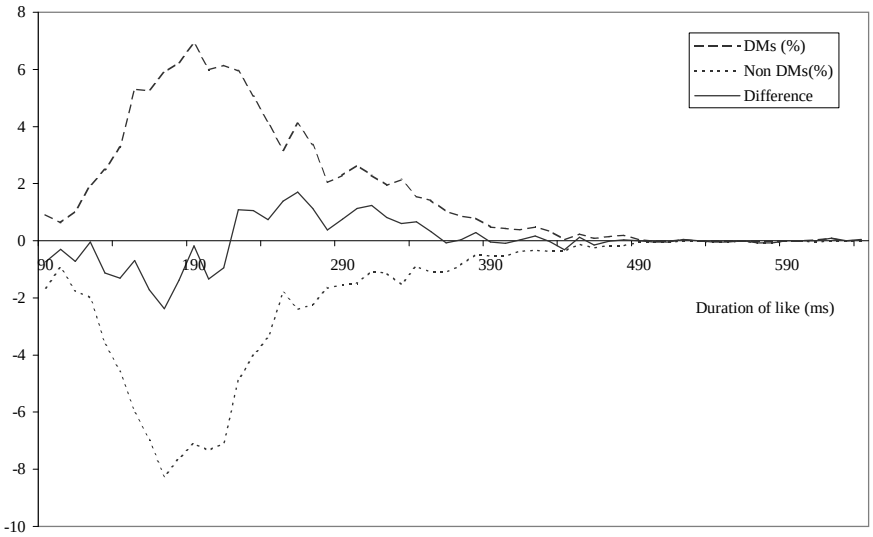

Figure 6.1: Duration of *like* as a DM (upper curve) vs. non DM (lower curve, reversed) as a percentage of all DM vs. non DM occurrences. The difference between the two curves (solid line) indicates for each duration whether there are proportionally more DM or more non DM occurrences of *like*.

deviations here show that the values vary considerably (see last lines of Table 6.3). In addition, the numbers of occurrences on which these values are computed are quite small: there only 884 DMs *well* which are not utterance initials (out of 3,639 DMs) and only 487 non DMs (in fact, only ten non DMs are utterance initials—six of them occur in the utterance: "Well done!").

The average durations of *like* and *well* appear to be quite close, as shown in Table 6.5. As with the previous figures, although the 95% confidence intervals are small, this does not mean that the durations of each class of occurrences are really concentrated around the average[5]. The standard deviations show in fact that the actual durations vary considerably with respect to the average value, though it remains true that on average, *like* as a DM is slightly longer than when it is not a DM; the opposite holds for *well*, which is shorter when it is a DM.

Given the variation of durations, it is of course not possible to distinguish DMs from non DMs on the basis of their duration, a fact appears even more clearly from the quantitative analysis expressed graphically in Figure 6.1 for *like* and Figure 6.2 for *well*. These graphs show the proportion of occurrences as a DM (upper curve) and as a non DM (lower curve, axis

[5]For instance, it is not true that 95% of all DMs *like* have a duration in the [222 ms; 228 ms] interval.

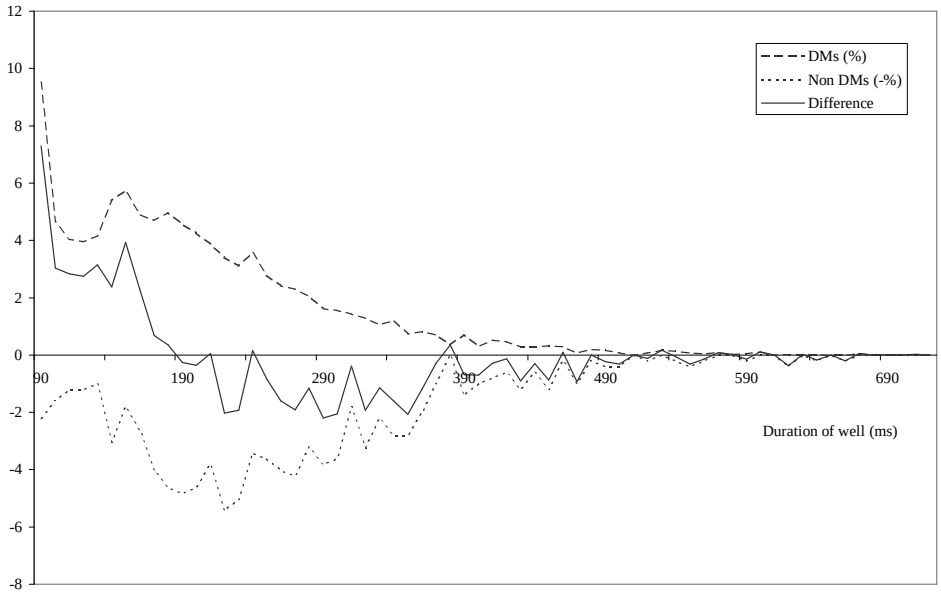

Figure 6.2: Duration of *well* as a DM (upper curve) vs. non DM (lower curve, reversed) as a percentage of all DM vs. non DM occurrences. The difference between the two curves (solid line) indicates for each duration whether there are proportionally more DM or more non DM occurrences of *well*.

reversed) for each observed duration[6]. For instance, in Figure 6.1, it appears that 5.9% of the occurrences of *like* as a DM have a duration of 180 ms, while about 8.3% of the occurrences of *like* as a non DM havethe same duration. Therefore, for *like* the are proportionnally more non DMs that last 180 ms than DMs. In other words, if an occurrence of *like* lasting 180 ms would have to be disambiguated with no other information available, it would be wiser to hypothesize that it is not a DM[7].

It appears that for *like*, there are three intervals of the duration which indicate a visible statistical difference between DMs and non DMs. Below ca. 220 ms, non DMs are in excesss, while between ca. 220 ms and 360 ms, there are more DMs than non DMs. For durations greater than 360 ms, there is no clear-cut difference. It appears that on the average, DM *like* is characterized by a smaller range of variation than non DMs. Furthermore, the existence of three different intervals may be due to the fact that as a non DM, *like* can fulfil a number of different functions, among which some may be less prosodically marked than the DM, and some others more (the latest appear to be far less frequent). More detailed labelling of the non DM functions are required to test this hypothesis.

In the case of *well*, the results (Figure 6.1) are less clear since the number of non DM occurrences is far smaller than the number of DMs. It appears that below 190 ms, the proportion of DMs exceeds that of non DMs, and for longer durations the two proportions are quite similar. Therefore, it appears that DM *well* tends to be shorter than non DMs.

[6]The curves are interpolations from durations expressed in tens of milliseconds

[7]This reasoning will be used when tying to disambiguate candidate DMs automatically, in particular in section 8.3.2 below.

6.2.2 Hypothesized Features

We propose therefore to use the following features, which strike a good balance between informativeness and tractability. We also hypothesize their roles in the identification of DMs, a role that will or not be confirmed by empirical studies.

initial: set to 'yes' if the token is the first word of an utterance, to 'no' otherwise. Hypothesized role: uncertain; may be an indicator for DM *well*.

final: set to 'yes-completed' if the token is the last word of a completed utterance, to 'yes-interrupted' if it is the last word of an interrupted utterance, and to 'no' otherwise. Hypothesized role: uncertain, probably token-dependent.

pause-before: duration (in seconds) of the pause before the token, or 10 seconds if the utterance starts with the token. Hypothesized role: a pause could indicate a DM.

pause-after: duration (in seconds) of the pause after the token, or 10 seconds if the token ends the utterance. Hypothesized role: a pause could indicate a DM.

duration : duration of the pronounced token itself. Hypothesized role: uncertain.

The segmentation of the data into utterances, as well as the indication of whether the utterance was completed or interrupted, are provided with the dialogue act annotation of the ICSI Meeting Corpus (Shriberg et al., 2004). We use here this ground truth segmentation while acknowledging that an automatic segmentation would not have quite the same accuracy (Stolcke and Shriberg, 1996) and would be thus responsible for a number of DM classification errors.

The timing of the tokens is obtained from the word-timing also provided with the resource. This was generated at ICSI using an ASR system, and is not fully accurate. More accurate information, especially more detailed prosodic features such as pitch variation, could be extracted from the sound files using a dedicated sound processing tool, but this operation seems difficult to automate entirely. There is a trade-off here between the difficulty to detect some prosodic aspects, and the risk that they are not used, after training, by the automated classifiers.

6.3 Sociolinguistic Features

Speaker-related information, when available, seems *a priori* relevant to DM identification. For instance, if a speaker or class of speakers appear to have a marked preference for DM-uses of *like*, this could help to disambiguate tokens left unsolved by the other features—something useful in a dialogue processing system that is used frequently by the same persons.

The real identity of the speakers is not available in the ICSI Meeting Corpus, but anonymous codes are used to uniquely identify them in transcripts, and some sociolinguistic information is provided (see Section 3.3 above and Appendix 3.2). We hypothesize the role of the following sociolinguistic features:

gender: 'female' or 'male'. Hypothesized role: no influence.

age: an integer. Hypothesized role: younger speakers have preference for DM *like*.

education: 'undergraduate', 'graduate', 'PhD', 'professor'. Hypothesized role: higher education entails less DM use.

native: 'native' vs. 'non-native' English speaker. Hypothesized role: no influence.

Feature	Value	Nb. of $like_{DM}$	$like_{DM}$ /words	$like_{DM}$ /$like_{all}$	Nb. of $well_{DM}$	$well_{DM}$ /words	$well_{DM}$ /$well_{all}$
Gender:	male	1,144	.19%	40%	2,754	.51%	88%
	female	908	.51%	55%	885	.56%	89%
English	native	1,586	.27%	44%	2,655	.52%	87%
language:	non native	466	.22%	49%	984	.52%	90%
Origin:	US East	281	.13%	31%	844	.46%	84%
	UK	32	.28%	40%	81	.83%	86%
	other US	785	.31%	46%	1,389	.61%	89%
	other c.	466	.22%	49%	984	.52%	90%
	US West	488	.50%	55%	341	.38%	91%
Education:	professor	169	.07%	22%	880	.47%	84%
	PhD	898	.29%	48%	1,722	*.61%	*90%
	graduate	799	.33%	50%	987	*.47%	*88%
	undergrad.	186	1.26%	67%	50	.36%	94%
Average	values	2,052	.26%	45.4%	3,639	.46%	88.0%
Standard	deviations	—	.52%	27.2%	—	.29%	30.2%

Table 6.6: Dependency of the frequencies of DMs *like* and *well* on speaker-related features. For each feature, classes are sorted by increasing frequency. This order appears to be the same for *like* and *well*, except in one case marked with a '*'.

origin: 'UK', 'US East', 'US West', 'US other', 'other'. Hypothesized role: uncertain; possibly speakers from US West have a preference for DM *like*.

The three types of features described in this section can generalize to other DMs and to other conditions of use. It is true that the lists of collocations must be built from scratch for each new DM, but this is not very time-consuming, and can be done using hints from the literature together with distributional studies, provided a sizeable amount of data is available to test the generality of collocations. Prosody, position and sociolinguistic features are not token dependent, though the way they are used may vary.

The analysis of the raw frequencies shows that the preference for *like* as a DM shows much greater variability than the preference for *well*, which seems more neutral. About 45% of all occurrences of *like* are DMs (27% standard deviation), whereas 88% of the occurrences of *well* are DMs, a much higher proportion. Relative to the total number of words, the frequency of *like* as DM is 0.26%, with 0.27% standard deviation (!), while the frequency of *well* is 0.46% with a similar standard deviation.

The frequencies do not vary significantly between native and non-native English speakers. However, there is interesting variation for the distribution of *like*, which seems to be preferred by speakers from the US West, as opposed to US East, and by lower-educated speakers. Indeed, speakers from US West use 0.50% DM *like*, other US regions use 0.31%, other countries 0.22%, US East 0.13% (the result holds also when counting DM uses among all uses of the token *like*, from 55% to 31%). Also, undergraduates use 1.26% DM *like*, graduates 0.33%, PhDs 0.29%, professors 0.07%. However, in the present recordings, speakers from US East happen also to be older and more educated than those from US West (see Appendix 3.2), so more speakers are needed to find out which of the two sociolinguistic factors, region of origin or education, influences most the use of *like*. As for *well*, speakers from US West seem to use it less often than those from US East or other countries, but these preferences are less marked than in the

case of *like*.

6.4 Dialogue Acts

An experiment was also conducted to study the relevance of features such as dialogue acts associated to each utterance by human annotators Shriberg et al. (2004). This reference annotation uses the ICSI Meeting Recorder Dialogue Act tagset, which allows for each utterance the use of complex dialogue act labels made of one first level tag, zero or more second level tags, and a disfluency marker indicating whether the utterance was interrupted or aborted (see (Popescu-Belis, 2005) for a discussion). In addition, some utterances are further segmented in relation to dialogue act annotation, in which case labels are separated by a '—' sign, or by a ':' in case of a reported speech (direct quotation).

Two methods to encode this feature were explored. The first one involves decomposing the dialogue act label into tags, and encoding the first one as a da_tag_1 variable, the second one as da_tag_2, etc. There are at most six tags per label in the data, and since the last one is used to signal disfluencies, it is encoded as da_tag_{last}. However, this encoding does not take correctly into account the '—' and ':' signs.

The second encoding that was studied consisted simply of a *da_tag* feature that encoded for each utterance the full label, on condition that the label's frequency in the corpus was greater that a certain threshold, typically 20 occurrences.

It is not clear that dialogue act labels are related to DM uses of *like* and *well* within utterances. We conducted an experiment to assess this hypothesis, but in most of the experiments the dialogue act features were not used.

6.5 Summary of Features

To summarize, there are two parameters that determine the nature of the features:

1. $2N$, the size of the lexical window around the candidate DM;
2. F, the cut-off frequency for collocated words, i.e. words that appear at least F times in the windows of size $2N$ around all candidate DMs. Let $\mathcal{W}$ be this set.

The full list of features with possible values is the following (the first two items are *alternative* codings of lexical features):

- **either** $word_{-N}$, ..., $word_{-1}$, $word_{+1}$, ..., $word_N$ — for each n, $word_n \in \mathcal{W} \cup$ {'absent', 'other'}
- **or** position($word_i$) — for each i, position($word_i$) $\in \{-N, -N+1, \ldots, 0, 1, \ldots, N-1, N\}$
- initial — 'yes' or 'no'
- final — 'yes-completed', 'yes-interrupted', 'no'
- pause-before — length in seconds
- pause-after — length in seconds
- duration — length in seconds
- gender — 'female', 'male'

- age — integer (years)
- education — ‘undergraduate’, ‘graduate’, ‘PhD’, ‘professor’
- native — ‘native’, ‘non-native’ (English speaker)
- origin — ‘UK’, ‘US East’, ‘US West’, ‘US other’, ‘other’
- token — *like* or *well*

Chapter 7

Machine Learning: Decision Trees and Other Classifiers

At this stage, the DM identification task has been identified as a classification task over the set of tokens that are potential DMs, here the tokens *like* and *well*. In addition, an annotated set of over 4,000 instances of each token was made available, and a number of features that appear to be correlated with the DM vs. non-DM distinction and that can be processed automatically have been identified. The choice of a classification algorithm that can be trained using machine learning will now be discussed (Section 7.1), as well as the use of the data for training vs. test (Section 7.2).

7.1 Types of Classifiers and Training Methods

Several machine learning methods are designed to construct classifiers based on observations of feature/class association on training data. In the case of DMs, some of the features are discrete while others are continuous; moreover, the lexical features are quite particular as they are sparse, and their (linguistic) relation to the classification problem is very complex. The volume of training data, although larger in this study than in most earlier ones, remains quite small with respect to the range of possible values of the features.

A number of statistical classification algorithms were considered, such as decision trees, Naive Bayes, support vector machines, Bayesian networks, or k-nearest neighbors. Hidden Markov models do not seem to be particularly relevant because the candidate DM instances are not consecutive in the dialogue, as they can be separated by one or more utterances that do not contain a candidate DM, so there is no direct influence from one instance to the following one[1].

7.1.1 C4.5 Decision Trees

We believe that decision tree classifiers, trained using machine learning, are one of the best tools for DM identification, for a number of theoretical reasons that also receive empirical confirmation in the next chapter.

To summarize briefly their principle, decision trees are made of nodes that represent tests on one feature of a DM-candidate and branches that stand for the possible values of the features. To each terminal node, or leaf, is associated one of the two classes, 'DM' or 'not DM'. In order

[1]Of course, the status of the previous utterance, for instance in terms of dialogue acts, may influence the identification of utterance-initial DMs *well*—a problem that we leave to future studies.

to decide whether a token is a DM or not, the classifier starts at the root node and performs successive tests on the features related to the token, until a leaf is reached and a decision is made.

Decision trees can be learned from training data using the well-known C4.5 method (Quinlan, 1993), which accepts a mixture of discrete and continuous features for each instance. The C4.5 method automatically constructs a nearly optimal decision tree classifier for the training data, that is, a tree that maximizes the number of correctly classified instances (CCIs) over the training data. It is important to note that the goal of C4.5 decision tree learning is to maximize the number of CCIs and not the recall or precision or *kappa* of DM identification. Of course, a high umber of CCIs over the training data does not *a priori* imply that the resulting classifier will also score high for unseen test data, hence the importance of objective evaluation.

The greatest advantage of the C4.5 method in the present case is that it provides an explicit, interpretable classifier, unlike support vector machines or neural networks. Other classifiers with the same property are discussed below, however, this study is focussed on decision trees as they are fast to train and provide easy to understand discrete classification rules.

Decision trees with C4.5 training were already used by Siegel and McKeown (1994) and Litman (1996) to identify DMs, though in the first paper none of the resulting trees improved the performance over the baseline. Decision trees were also used by Heeman and Allen (1999) to estimate the probabilities of their POS-based language model including the DM tags. For other tasks related to discourse processing, decision trees were used for utterance segmentation (Passonneau and Litman, 1997) and dialogue act tagging (Shriberg et al., 1998).

7.1.2 Adjusting the C4.5 Decision Tree Learner

One of the properties of C4.5 decision trees is that the higher a feature appears in the resulting tree, the more general and discriminative it is. The accuracy of each leaf—the proportion of tokens correctly classified by the conjunction of the tests leading to that leaf—provides explicit information about the relevance of each feature. In particular, if a feature does not appear at all in the classifier, then it is not enough correlated with the DM/non-DM distinction, or it is superseded by other, more discriminative features.

Several parameters can be tuned to improve the accuracy of the C4.5 decision trees. The most significant ones are the possibility to search for binary decision trees, i.e. with exactly two branches for each test node, and the possibility to require that each leaf classifies at least N tokens (e.g. $N = 5$ or $N = 20$) from the training data—a constraint that is meant to preserve the generality of the resulting decision tree, i.e. to avoid over-fitting the training data and the risk of much poorer performance on test data. In addition, smaller values lead to large trees, which are more difficult to analyze.

In this study, we used the implementation of C4.5 and of the other classifiers from the Waikato Environment for Knowledge Analysis (WEKA), a helpful machine learning toolkit designed by Witten and Frank (2000) and made available by the authors[2].

7.1.3 Other Classifiers

We also explored the performances of Naive Bayes and Bayesian Network classifiers, Support Vector Machines (SVMs) and k-nearest neighbours, all in the WEKA implementation, mainly for empirical comparison purposes. SVMs are known to be powerful, non-linear (depending on the kernel) classifiers, but they appeared to require huge computational resources in terms of

[2]Open source software at: `http://www.cs.waikato.ac.nz/ml/weka/`.

memory and training time. In the few cases that were tested—on single training/test folds—their performances were not higher than those of the best decision trees. The length of the training prevented us from exploring various kernel functions.

The k-nearest neighbors method, with $k = 3$ and the full training set acting as a classifier, was tested as a baseline classifier along with the majority classifier (WEKA's 'ZeroR'), which classifies all instances as the most frequent class. The method does not require particular training, neither does it provide explicit classification rules.

The Naive Bayes classifier did not provide encouraging results: though fast to train, it is likely that the DM identification problem was too complex for this type of classifiers due to the lexical features. However, the Bayesian Network classifier, though much longer to train, provided some of the best scores that were obtained. These scores were only slightly higher than those of decision trees, and given that the probabilistic network is slightly less interpretable than decision tree nodes, this direction was not much pursued yet.

7.2 Use of the Data for Training and Test

For testing, we experimented first with separate training and test sets derived from the data, and then by using ten-fold cross-validation of classifiers. This procedure, which enabled us to make efficient use of all the available data, divides the data into 90% for training and 10% for test, repeating this process ten times, and computing the average score obtained on test data, along with confidence intervals.

While the WEKA interface offers a built-in capability for n-fold cross-validation, its results do not include confidence intervals or other measures of the variance. Therefore, these figures had to be computed outside the WEKA interface, using Java calls to the WEKA functions that build, store and evaluate classifiers.

The WEKA functions were also helpful to prepare the training and test data. The data was stratified according to the DM / non DM classes, i.e. the training and test sets that were generated contained the same proportion of DMs vs. non DMs, though not necessary the same proportion of *like* vs. *well*. For additional experiments focussing on each token, separate sets of ten training and test pairs for *like* and respectively for *well* were also created, still stratified for DMs vs. non DMs.

The separation of 90% training and 10% test data can be done either with randomized folds, or simply in the order of occurrence of the candidate DMs, which is the order of the concatenated meetings of the ICSI-MR corpus—both have been tested. In principle, randomized folds make data more homogeneous, which should reduce standard deviation and confidence intervals. Conversely, using non-randomized folds amounts to more strict evaluation conditions, as the variability between the training and test data increases, e.g. in terms of speaker participation, as speakers are not evenly distributed across the 75 meetings.

Therefore, non-randomized folds are used in the experiments that follow, to avoid using a lenient, favorable evaluation procedure. A number of trials comparing randomized vs. non-randomized folds have shown indeed that performances are slightly higher when using randomized folds—as the training data is more similar to the test data—but unlike what was expected the confidence intervals were not smaller.

The comparison of various classifiers has to take into consideration the confidence intervals computed by ten-fold cross-validation, as Litman (1996, section 3.3), for example, clearly explains:

> To determine whether the fact that an error rate E_1 is lower than another error rate E_2 is also significant, statistical inference is used. In particular, confidence intervals

> for the two error rates are computed, at a 95% confidence level. When an error rate is estimated using only a single error rate on a test set (i.e., the train-and-test methodology), the confidence interval is computed using a normal approximation to the binomial distribution [...]. When the error rate is estimated using the average from multiple error rates (i.e., the cross-validation methodology), the confidence interval is computed using a t-Table [...]. If the upper bound of the 95% confidence interval for E_1 is lower than the lower bound of the 95% confidence interval for the error rate E_2, then the difference between E1 and E2 is assumed to be significant.

Therefore, confidence intervals at 95% level were computed using Student's law, which is precisely the one used in the t-test tables in the quote above. This law is stricter than the normal law which is sometimes used to compute confidence intervals, which means that intervals tend to be larger (which is less desirable) when computed using Student's law. These figures allow, in what follows, an assessment of the significance of numerical differences between scores in various experimental conditions: if two confidence intervals are disjoint, then one classifier is better than the other with 95% confidence. If they are not, then the two classifiers have comparable performances.

In the remainder of the study, classifiers constructed with various learning conditions and on various feature sets will be compared in order to find the best DM disambiguation performances and the most relevant features for this task.

Chapter 8

Automatic Identification of DMs: Results

This section discusses the results obtained with machine learning classifiers on the DM identification task. First, empirical evidence is brought regarding the best settings for recognition experiments in terms of classification methods (Section 8.1). Then, the variations of the scores are studied, that is, the lowest (baseline) scores (Section 8.2.1) and the best obtained results (Section 8.2.2). The relevance of each type of features to the classification problem is then analyzed, using in particular experiments when the type of features is removed or, conversely, used independently (Section 8.3). A more systematic analysis of the features is then done using automatic attribute selection techniques (Section 8.4). The final section summarizes these findings and compares them with previous results (Section 8.5).

8.1 Comparison of Machine Learning Methods

Apart from the theoretical reasons that motivate the choice of a machine learning method to train a DM classifier[1], the empirical results obtained with various methods implemented in Weka also shed light on the quality of each method.

Experiments with C4.5 decision tree learners, support vector machines (SVMs), k-nearest neighbors and Bayesian networks applied to DM identification suggest that sometimes the best scores reached by different methods do not differ substantially. In particular, the differences are somewhat smaller than the differences between the results of different configurations of a same method. In any case, all these scores have to be compared with baseline scores obtained by trivial classifiers, as shown in the next section.

A number of tests were done to compare these methods, by keeping some parameters constant and changing other parameters one by one. One of the main options is the encoding of lexical features: for instance, when using a lexical window of size one (one word before and one after the candidate DM) and considering as possible values of the features WORD(-1) and WORD(+1) all the words appearing at least three times in this window, the C4.5 decision tree reaches $\kappa = 0.723 \pm .009$ while the 3-nearest neighbours classifier reaches only $\kappa = 0.681 \pm .014$. This type of comparisons is the basis for the following findings.

As shown in Tables 8.4 and 8.5 of the next section, the overall best scores are obtained by a Bayesian network classifier that uses only discrete features, including the dialogue acts (the most frequent labels), but excluding all the duration dependent features. The scores of this

[1] For instance, the size of the search space, the number of training examples, the number and nature of available features—as discussed for dialogue acts in (Popescu-Belis, 2005, section 6).

classifier are slightly higher than those of the best SVMs and C4.5 decision trees, the difference being significant in terms of 95% confidence intervals or paired t-tests.

The Bayesian network classifiers take however longer to build and are less interpretable, in the Weka environment, than C4.5 decision trees. Both are however considerably faster than SVMs [2]). The C4.5 method will be preferentially used in most of the experiments below, due to the understandability of the decision trees and the efficiency of the training.

The k-nearest neighbours classifiers lead to significantly lower results than the other, non-baseline methods. It appears that $k = 3$ represents a good compromise between performance and classification speed, as well as the use of around 200 DM instances for classification. The scores of this method are still well above baseline scores, as shown in Tables 8.4 and 8.5.

Similar analyses help finding the best parameters for building C4.5 decision trees. One of the most important parameters is the minimum number of instances from the training data categorized at each leaf of the decision tree[3]. A comparison was conducted using the WORD(-1) and WORD(+1) features, having as possible values all the words appearing at least three times in this [-1; +1] window. Comparing classes of, respectively, at least 2, 10 and 50 instances, the scores of the classifier are, respectively, $\kappa = 0.733 \pm 0.010$, $\kappa = 0.723 \pm 0.013$ and $\kappa = 0.716 \pm 0.015$. As expected, when smaller classes are allowed, the accuracy of the classifier increases, but the differences are not significant at the 95% confidence level (the confidence intervals are not disjoint). Given the variety of candidate DM instances, a minimum value of 10 instances at each leaf seems a good compromise between accuracy and intelligibility of the resulting classifier.

Other comparisons indicate for instance the best ways to encode some of the features used for DM identification, in particular lexical ones, as shown below in Section 8.3.1.

8.2 Comparison of Lowest and Highest Scores

To assess the significance of the best DM identification scores, it is necessary to compare them with baseline ones, that is, with the performances of the simple classifiers that score well above zero. For instance, as mentioned in Section 2.3 above, in the experiments conducted by Siegel and McKeown (1994), the elementary classifier using only the following rule: "if the candidate DM is preceded by a comma or a period (i.e. if it is utterance-initial) then it is a DM, otherwise it is not", reached 79.16% accuracy, a score that was never significantly surpassed in that study by more complex, better trained classifiers. This simple rule could therefore be considered as an (informed) baseline score.

We will start by presenting a series of baseline scores for a number of elementary classifiers, and then present the best scores we obtained. More detailed discussions about how these scores were obtained and the utility of the various types of features will be presented in the subsequent sections.

8.2.1 Baseline Scores

The five metrics that we use for scoring DM identification vary between 0 and 1, with 1 being the best score. In fact, κ varies from -1 to $+1$; negative values signal an opposite correlation between the two annotations, and 0 signals random correlation. However, the classifiers produced by Weka never have negative κ scores because it is sufficient, in such a situation, to

[2]The construction of one SVM takes more than four hours on a SunBlade 100 workstation with a 400 MHz processor and 512 MB memory.

[3]As mentioned in the previous section, trees allowed to use smaller classes are more accurate but are also more complex and harder to interpret, and may lack generality on other data sets.

Training	Test	Class	CCIs (%)	κ	Recall	Precision	F-measure
like+well	*like+well*	DM	65.754±.030	0	1	.658±.0004	.794±.0004
like+well	*like*	DM	45.397±1.098	0	1	.454±.011	.624±.010
like+well	*well*	DM	87.989±1.163	0	1	.880±.012	.936±.007
like	*like*	non DM	54.592±.063	0	0	0	0
well	*well*	DM	87.984±.071	0	1	.880±.001	.936±.0003
like+well	*like+well*	*like:* non DM *well:* DM	70.549	.419	.639	.880	.741

Table 8.1: Baseline scores obtained by a majority classifier. The majority class assigned to all instances is shown in the third column, for each of the experimental conditions indicated in the first two columns. Decimals and confidence intervals are not reported when theoretical reasons show that no variation occurs across folds (i.e. '0.000±0.000' is written '0'). The last line indicates the scores of a token-specific majority classifier.

reverse the binary classification in order to obtain positive correlation[4].

Null scores are seldom observed for some of the metrics, because of the binary nature of the classification, which leads to a random performance of 50% (asymptotically). As shown in Table 8.1, the *majority classifier*, which assigns to all candidate DMs the type of the most frequent class observed in the training data—which is almost always 'DM' except when restricted to *like*—reaches scores that are well above zero for at least three metrics out of five. Only κ appears to be insensitive to this bias and exhibits null values for the majority classifier.

Table 8.1 also illustrates five experimental conditions that will be often used for testing. The first condition uses ten-fold cross-validation on the whole data set, that is, training and testing the classifiers on both lexical items at the same time. The second and third conditions display the scores of a classifier trained on both *like* and *well* but tested only on, respectively, *like* then *well*. So, this is in fact the same classifier as in the first condition, with scores for the identification of DM *like* vs. *well* made visible. In the fourth and fifth conditions, classifiers are trained and tested separately on each type of candidate DM. This type of experiments aims at finding out whether one DM is more difficult to identify than the other, and whether the useful features are different or not.

The majority classifier, when trained and tested on the same token, necessarily classifies correctly at least half of the instances, i.e. CCI $\geq$ 50%, as it appears in lines 1, 4 and 5 from Table 8.1). However, $\kappa = 0$ in all five conditions since the κ score factors out the probability that a candidate is classified as a DM by chance, which is equal t the ratio of DMs in the data. Because recall and precision are defined, asymmetrically, only for the DM identification task, their values are well above zero when the majority classifier assigns the 'DM' label to all instances (which is the case in all conditions but one): recall is 100% as all correct DMs are identified, while the precision is the ratio of non DMs among all tokens. Therefore, a number of very high values appear for f-measure, a fact that must be kept in mind for future comparisons.

The confidence intervals are also much smaller (at least ten times narrower) for the majority classifier in the first, fourth and fifth condition than in the other two. In fact, a number of metrics such as κ and recall remain constant through the ten-fold training/test procedure, which is signalled in Table 8.1 by removing the indication of the confidence interval (instead of writing ±0). There is however some variation in the number of CCIs across the folds, hence nonzero confidence intervals, due to the stratification procedure used to construct the folds. For

[4]In other words, it is always possible to have $\kappa \geq 0$, i.e. to have more correctly classified instances than incorrectly classified ones, for a binary task.

instance, in the second and third conditions, the folds are stratified according to the proportion of DMs vs. non DMs[5], but the proportions of *like* DM vs. non DM, and respectively *well* DM vs. non DM are not necessarily the same in each fold.

Other simple classifiers could be used as a baseline. The performances of some of them will be reported below, in relation to the subset of features they use. It is important to note that using only the TOKEN feature, i.e. allowing a token-specific majority classifier to distinguish between the tokens *like* and *well*, already increases the scores to higher values than the majority classifier. For instance, the decision tree classifier based only on the following rules: "*like* is not a DM" and "*well* is a DM" reaches the following scores: CCI $= 70.5488\%, \kappa = 0.419, r = 0.639, p = 0.880, f = 0.741$. Although the f-measure is slightly lower than for the majority classifier shown in the first line of Table 8.1, κ is considerably higher; this might as well reflect a possible bias of κ towards high values for very weak performances.

8.2.2 Highest Scores

The use of all possible features enables three machine learning methods to reach similar best classification accuracies, with a slight advantage to the Bayesian network classifier. The best results of this classifier, which are also the best overall scores obtained in our experiments, are shown in the first line of Tables 8.2 and 8.3 (the latter table offers more detailed figures). As in the previous tables, the scores of this classifier restricted to *like* and respectively *well* are shown in the second and third lines, and the scores of the same learning method trained and tested separately on *like* and *well* are shown in the fourth and fifth lines.

The best scores are significantly above the baseline. The fact that $\kappa = 0.78$ shows that the DM classification performance is in the same range as human inter-annotator agreement. The value of κ is significantly higher (with 95% confidence) than any of the baseline classifiers described above, as the best one from Section 8.2.1 had $\kappa = 0.42$. The best scores are also higher than most of the classifiers that use only a subset of the features, described hereafter. Of course, scores such as CCI $= 90.5\%$ and $f = 0.93$ are also clearly higher than those obtained by the baseline methods. The best scores will be compared to those obtained by other studies in Section 8.5.3 below.

The performances of the best classifier are significantly higher for the identification of DM *well* ($f = 0.986$) than for *like* ($f = 0.836$). The DM *well* appears thus to be easier to identify than *like*, at least given the feature set used here. It is also true that *well* as a DM is much more frequent than *like* as a DM (ca. 88% vs. 45%[6]) but the effect of DM frequency on the scores should be filtered out at least by the κ metric, which is not the case: for *well*, $\kappa = 0.88$ is also much higher than $\kappa = 0.68$ for *like*. Similar DM-specific scores are obtained when the Bayesian network classifier is trained only on instances of *like*, respectively *well*, as shown in the fourth and fifth lines of Table 8.2.

A comparison of the best scores obtained by different classifiers appears in Tables 8.4 and 8.5. The best scores of the SVMs or C4.5 decision trees are quite similar to those of the Bayesian network. In fact, even though these scores appear to be slightly inferior, the comparison is not significant at 95% because the confidence intervals are overlapping, except only for κ.

Conversely, all the three methods generate significantly better classifiers (with 95% confidence) than the best k-nearest neighbours classifier (which is $k = 3$), whose scores appear in the fourth line of Tables 8.4 and 8.5. The 3-nearest neighbour classifier is in turn significantly better than the baseline majority classifier shown in the fifth line. The five evaluation metrics

[5]That is, each training set contains 90% of all instances with the same proportion of DMs vs. non DMs as in the whole data set.

[6]See Table 6.6 of Section 6.3 and also Section 3.3.

Training	Test	CCIs (%)	κ	Recall	Precision	F-measure
like+well	*like+well*	90.5	0.78	0.96	0.90	0.93
like+well	*like*	84.0	0.68	0.90	0.78	0.84
like+well	*well*	97.5	0.88	0.99	0.98	0.99
like	*like*	84.5	0.69	0.90	0.79	0.84
well	*well*	97.5	0.88	0.99	0.98	0.99

Table 8.2: Best overall results for training and test on both DMs (first line), obtained by a Bayesian network classifier. Its scores restricted to *like* and respectively *well* appear in the second and third lines. The fourth and fifth lines show the performance of Bayesian network classifiers trained and tested respectively on *like* and *well*. More decimals and confidence intervals are shown in Table 8.3 below.

Training	Test	CCIs (%)	κ	Recall	Precision	F-measure
like+well	*like+well*	90.480±.646	.783±.016	.957±.004	.904±.008	.930±.005
like+well	*like*	84.009±1.431	.681±.028	.896±.012	.784±.021	.836±.014
like+well	*well*	97.537±.456	.880±.021	.991±.004	.981±.005	.986±.003
like	*like*	84.510±.983	.691±.020	.899±.015	.789±.008	.840±.011
well	*well*	97.510±.489	.877±.024	.993±.003	.979±.004	.986±.003

Table 8.3: Best overall results, obtained by a Bayesian network—same as Table 8.2 but with confidence intervals and three decimals instead of two.

are in most cases concordant with respect to the ranking of the various methods.

The best configuration for the C4.5 decision tree learner encodes the lexical features as four variables (WORD(-2), WORD(-1), WORD(+1) and WORD(+2)) whose possible values are all the words appearing at least 10 times in this lexical window of size four. In this case, the C4.5 learner was set to construct binary unpruned trees only, with at least two instances per leaf, a parameter that increases precision but makes the resulting tree quite complex. A much simplified binary tree, reaching only $\kappa = 0.65$, is shown in Figure 8.1: some of the most discriminative values of the lexical features appear as upper branches, along with the 0.07 s limit of PAUSE-BEFORE and a similar value for PAUSE-AFTER. At the bottom, some sociolinguistic features are used for *like*. The figures show for each leaf the number of tokens correctly vs. incorrectly classified.

The average precision of the C4.5 decision trees is significantly higher (with 95% confidence) than the precision of SVMs and k-nearest neighbours, and is in the same range as the precision of the Bayesian network (it is even apparently higher, but the difference is not significant with 95% confidence).

8.3 Contribution of Features to DM Identification

This section examines the relevance of each type of feature described in Chapter 6. Essentially two types of experiments are conducted.

1. Classifiers are trained using only a specific feature and their performance is evaluated with respect to the baseline: the higher the difference, the more important the feature.

2. Classifiers are trained using all available features except a specific one, while their performance is evaluated with respect to the best overall scores: again, the higher the difference,

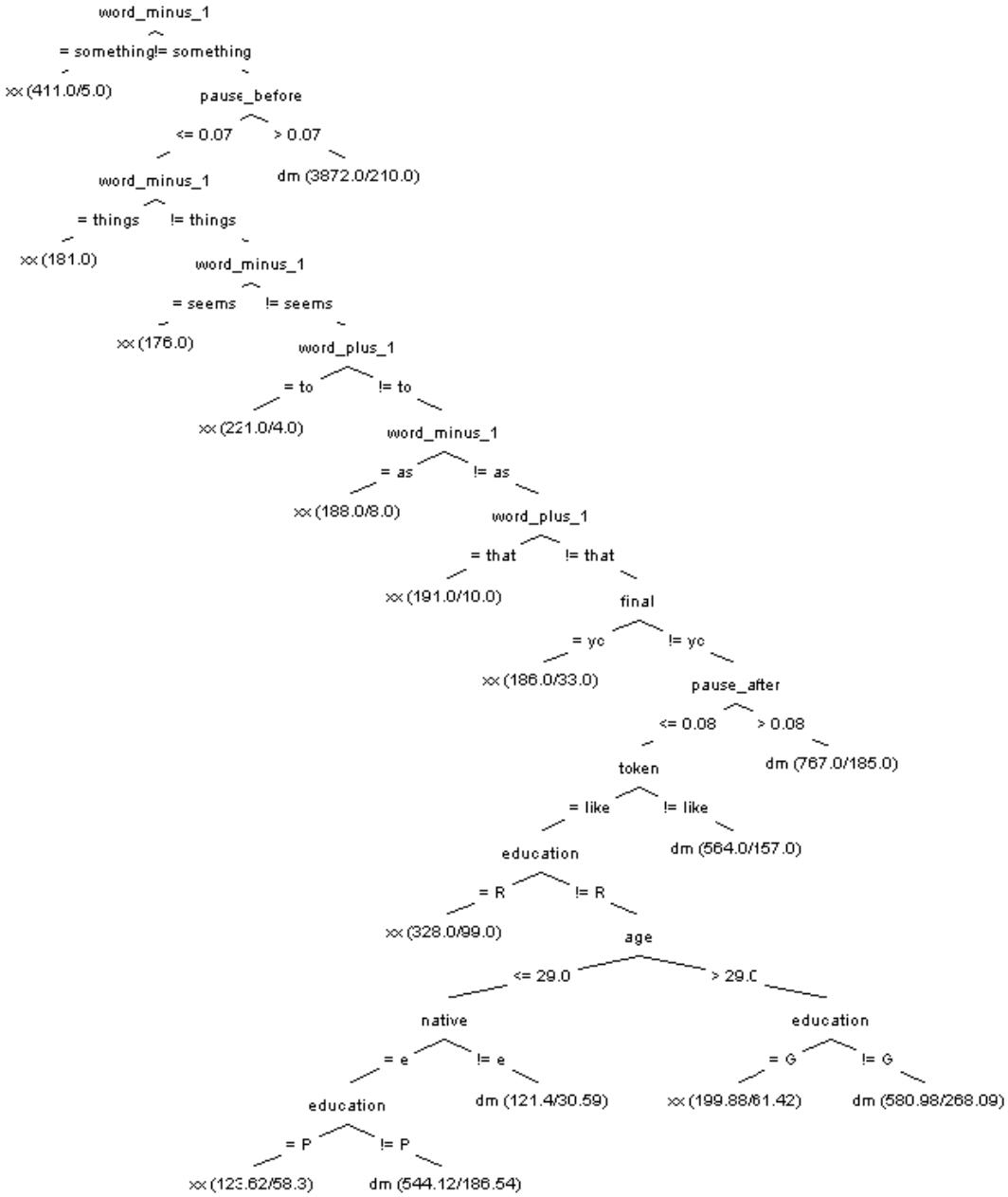

Figure 8.1: Simplified C4.5 decision tree using most features and scoring $\kappa = 0.65$.

Method	CCIs (%)	κ	Recall	Precision	F-measure
Bayesian network	90.5	0.78	0.96	0.90	0.93
SVM	89.3	0.75	0.96	0.88	0.92
C4.5	88.9	0.75	0.92	0.91	0.92
3 Nearest N.	85.5	0.68	0.88	0.89	0.89
Majority ('DM')	65.7	0	0.66	1	0.79

Table 8.4: Best results obtained by four machine learning methods in comparison with the baseline majority classifier. The ranking of the methods is fully concordant between the number of CCIs, κ and f-measure scores. More decimals and confidence intervals are shown in Table 8.5 below.

Method	CCIs (%)	κ	Recall	Precision	F-measure
Bayesian network	90.480±.646	.783±.016	.957±.004	.904±.008	.930±.005
SVM	89.290±.571	.752±.014	.964±.006	.884±.008	.922±.004
C4.5	88.862±.511	.751±.011	.923±.007	.909±.006	.916±.004
3 Nearest N.	85.546±.624	.681±.014	.885±.008	.894±.006	.890±.005
Majority ('DM')	65.754±.030	0	.658±.0004	1	.794±.0004

Table 8.5: Best results obtained by four machine learning methods and a baseline classifier—same scores as in Table 8.4 but with confidence intervals and three decimals instead of two. The sizes of the 95% confidence intervals do not allow a clear separation of the first three methods.

the more important the feature.

In the second case, even if the difference observed when removing a certain feature is small, this does not mean that this feature is totally irrelevant to the DM classification problem. It may just be the case that the feature is redundant when features with more discriminative power are available, but may still be relevant when such features are removed.

Such experiments should ideally be carried out using sets of independent features; otherwise it is not possible to fully remove a certain type of features. Here, positional and prosodic features are not independent, because the length of the pause before or after a DM candidate is necessarily an indicator of its utterance-initial or utterance-final character, as a specific value must be assigned to the temporal features. Similarly, the lexical features are related to the positional ones, because if the token is initial, the words before it (e.g. WORD(-1) or WORD(-2)) must be set to a specific value such as 'none'. This overlap between features makes conclusions about feature relevance less specific.

The analysis of feature relevance using attribute selection algorithms, performed in Section 8.4 below, provides more systematic results on the same problem.

8.3.1 Lexical Features

Encoding of Lexical Features

Two options are possible for the encoding of the lexical features, described in Section 6.1.3:

1. position-related variables, whose possible values are the words appearing above a certain frequency;

2. one variable per word, for words above a certain frequency, possible values being the positions of the word with respect to the candidate DM (-1, $+1$, etc.)

When comparing results for the C4.5 decision tree learner, all other parameters being kept constant, it appears that the scores for both options are quite similar, regardless also of the cut-off frequency (3 or 10). The lack of significant differences confirms the fact that the two options encode exactly the same information. The first option (positional variables WORD(−1),WORD(+1), etc.) will be used in what follows as it makes decision trees simpler and easier to interpret.

Size of Lexical Window

Empirical analysis also shows that a lexical window larger than $[-1;+1]$ around the candidate DM does not improve significantly the performance of DM identification. Therefore, only the WORD(−1) and WORD(+1) features will be used below.

Cut-off Frequency

The frequency threshold F determining the acceptable values of WORD(−1) and WORD(+1) was also studied empirically, i.e. the number of times a word must appear around a candidate DM in the whole corpus in order to be permanently considered as a lexical indicator. Section 6.1.3 above showed that 360 different words appeared more than 3 times, only 150 more than 10 times, and only 90 words more than 20 times. If the same statistics are made using a larger lexical window, $[-2;+2]$, then there are respectively 250, 160 and 70 words appearing more than 3, 10 and 20 times.

In principle, setting a low F increases the precision of the classifier on the training data, but might make less able to generalize on test data. Conversely, setting a higher F leads to simpler decision trees, better generalizations, but smaller precision. Experiments with $F = 3$, 10 and 20 show that there is no significant difference in performance (with 95% confidence), therefore F=10 will be used in what follows[7].

Individual Lexical Features

Turning now to experiments with partial feature sets, three experiments are particularly eloquent (as they were run with Weka's ten-fold validation procedure, no confidence intervals are available):

1. Using only the WORD(−1) feature (i.e. the lexical item preceding the candidate DM), C4.5 constructs trees that contain as the uppermost leaf the lexical collocations that are the most reliable indicators of a DM (more details in Section 8.4 below), including 'none' as an indicator of initial position. The ten-fold training/test procedure applied by Weka yields the following average scores: CCI $= 86.5\%, \kappa = 0.68, r = 0.97, p = 0.85, f = 0.90$, which are not much below the best possible ones.

2. When distinguishing *like* from *well* in the decision trees, thanks to the TOKEN feature in addition to WORD(−1), the scores increase slightly to CCI $= 87.4\%, \kappa = 0.72, r = 0.91, p = 0.90, f = 0.90$. The increase of κ is the most remarkable one, as the κ value observed here is quite close to the best one from Table 8.4. As with the other experiments of this series, the resulting tree is not particularly intelligible as there are 150 possible values in the tree (words in the $[-1;+1]$ window with $F \geq 10$). The uppermost branches contain the collocations described in the feature selection study below (Section 8.4).

[7]It was also observed that the confidence intervals are somewhat larger for $F \geq 10$ than for $F \geq 3$, which is not what was expected: we expected more variability in performance for the slightly over-adapted trees produced for $F = 3$.

3. Words situated after the candidate DM appear to be much less informative with respect to the DM vs. non DM distinction. If only TOKEN and WORD(+1) are used for training, the scores reach clearly lower values than the previous ones: CCI = 77.8%, $\kappa = 0.47, r = 0.91, p = 0.79, f = 0.84$.

4. Finally, when using all lexical features encoded as WORD(n) with a window of size four (i.e. TOKEN, WORD(−2), WORD(−1), WORD(+1), WORD(+2)), the results are comparable with those obtained only with the preceding word (point 2. above), except that recall increases considerably while precision decreases: CCI = 88.1%, $\kappa = 0.73, r = 0.93, p = 0.90, f = 0.91$.

The lexical features appear thus to be nearly sufficient for DM identification, as the scores obtained using these features are very close to the best overall scores. In fact, it is the word before the token (WORD(−1)) which is clearly the most important one for DM identification. This is a matter of interesting linguistic discussion, which is beyond our present scope.

The second experiment also shows that the best lexical indicators are not the same for *like* and for *well*: the commonalities between the two include only the test showing whether the candidate DM is initial or not, WORD(−1) = 'none'.

The entire list of most relevant lexical indicators appears in Section 8.4 below, which also demonstrates that the most relevant features are indeed the lexical ones.

8.3.2 Prosody and Position for DM Identification

A set of experiments was designed to test prosodic and positional features. As already noted, information about the positions of candidate DMs within the utterance is related both to lexical features—e.g. WORD(−1) = 'none' signals an utterance-initial candidate—and to the prosodic/temporal features. These features must indeed code the fact that a DM candidate is utterance-initial or utterance-final because a specific value must be assigned for the length of the pause surrounding it.

In the following experiments, decision tree classifiers were built using one, two or three features (positional and prosodic/temporal). Only the most interesting combinations of features are discussed, and the scores of the respective classifiers are summarized in Table 8.6 below.

Positional Features

When using only the INITIAL feature (possible values: 'yes' or 'no'), the resulting classifier scores above baseline in terms of κ and precision: CCI = 68.8%, $\kappa = 0.42, r = 0.54, p = 0.97, f = 0.70$. This classifier is of course particularly simple, consisting of the following rule: "if the candidate is utterance-initial, it is a DM, otherwise it is not a DM". The high value of precision in this case shows that tokens in initial position are very likely to be DMs. The low value of recall shows that there are also many other DMs that are not initial.

If the classifiers are allowed to use the TOKEN feature in addition to INITIAL, the optimal classifier found by C4.5 is slightly more complex than above, as it consists of the following rules: "for *well*, all candidates are DMs; for *like*, initial candidates are DMs, the others are not DMs"[8]. The scores of this classifier are: CCI = 73.4%, $\kappa = 0.46, r = 0.70, p = 0.87, f = 0.78$. These values are higher than the previous ones, which is quite expected since an additional feature was available. The only decrease is precision, probably because C4.5 optimizes CCIs, not precision or recall in particular.

[8]This can be expressed alternatively as: "if the candidate is initial, it is a DM; otherwise, if the candidate is *like*, it is not a DM, and if it is *well*, then it is a DM".

Position						Prosody					
Features	CCI	κ	r	p	f	Features	CCI	κ	r	p	f
T	70.5	0.42	0.64	0.88	0.74	T	70.5	0.42	0.64	0.88	0.74
I	68.8	0.42	0.54	0.97	0.70	B	74.2	0.50	0.65	0.94	0.77
T+I	73.4	0.46	0.70	0.87	0.78	T+B	75.3	0.48	0.75	0.86	0.80
F	67.5	0.09	0.98	0.67	0.80	A	67.5	0.09	0.98	0.67	0.80
T+F	72.5	0.46	0.64	0.91	0.75	T+A	75.8	0.50	0.74	0.87	0.80
T+I+F	75.8	0.51	0.71	0.90	0.79	T+A+B	79.4	0.55	0.82	0.86	0.84

Table 8.6: Results obtained by a C4.5 decision tree learner using various combinations of positional and prosodic/temporal features. On average, classification is improved as more features become available among the following ones: T: TOKEN, I: INITIAL, F: FINAL, B: PAUSE-BEFORE, A: PAUSE-AFTER). The temporal features incorporate information about position and therefore generally lead to superior results. Abbreviations r, p and f stand for, respectively, recall, precision and f-measure of DM identification; CCIs are expressed as a percentage of all instances.

An experiment with the FINAL feature leads to poor results, very close to the baseline: CCI $= 67.5\%, \kappa = 0.09, r = 0.98, p = 0.67, f = 0.80$. This is an empirical confirmation of the fact, already mentioned above, that the items following candidate DMs are not very relevant to DM identification.

The use of the TOKEN feature along with the FINAL feature quite naturally increases the performance in terms of κ, as the use of TOKEN alone considerably increases κ, as shown in Section 8.2.1 above and in the first line of Table 8.6. However, the increase in CCIs and f-measure is only moderate: CCI $= 72.5\%, \kappa = 0.46, r = 0.64, p = 0.91, f = 0.75$. The optimal decision tree consists of the following rules: "*like* is a DM only if it ends an interrupted utterance; *well* is not a DM only if it ends a completed utterance".

Finally, using the three features TOKEN, INITIAL and FINAL altogether, the score of the optimal classifier constructed by C4.5 is: CCI $= 75.8\%, \kappa = 0.51, r = 0.71, p = 0.90, f = 0.79$. These scores clearly improve on all the previous ones, at least in terms of CCIs, κ and f-measure.

Prosodic or Temporal Features

Turning now to the prosodic/temporal features, a similar series of experiments was conducted using the PAUSE-BEFORE and PAUSE-AFTER features, which in reality subsume the previous features as they implicitly code the position of the candidate DM. Therefore, scores in this series are expected to be slightly higher than the corresponding scores in the previous series, as shown comparatively in Table 8.6. The criteria found automatically by the C4.5 decision tree learner often correspond to the observations regarding the prosodic/temporal behaviour of DMs vs. non DMs made in Section 6.2 above.

Using only the PAUSE-BEFORE feature, the best scores are CCI $= 74.2\%, \kappa = 0.50, r = 0.65, p = 0.94, f = 0.77$. The corresponding decision tree consists of the following rules: "if the pause before the candidate is longer than 30 ms (including candidates in initial position), then it is a DM, otherwise it is not". The value of κ is significantly above the baseline obtained using only the TOKEN token feature.

With TOKEN and PAUSE-BEFORE, classification is slightly improved over the previous classifier in terms of CCIs and f-measure but not of κ[9]: CCI $= 75.3\%, \kappa = 0.48, r = 0.75, p = 0.86, f = 0.80$. These scores are also slightly higher than those obtained using the INITIAL

[9]As observed above, this is probably because the C4.5 decision trees are optimized for CCIs and not κ.

feature. The corresponding decision tree contains a finer-grained distinction than the one using only INITIAL: "*well* is always a DM; *like* is a DM only when the pause before is longer than 0.06 s". The 60 millisecond limit—which appeared already in our previous study (Zufferey and Popescu-Belis, 2004)—helps to identify DMs through their specific prosodic contour; conversely, non DM uses are in general not separated prosodically from the word preceding them. The 60 millisecond value also appears to be the average value of the duration of pauses before *like* when it is used as a DM, and an acceptable discriminative value between the averages of the pauses before *well* when it is a DM (ca. 300 ms) and when it is not (ca. 20 ms), as shown in Section 6.2 above and especially in Table 6.3.

For the PAUSE-AFTER feature alone, the tree constructed by C4.5 is exactly equivalent to the one obtained with the FINAL feature, and consists of the rule: "if the candidate ends a completed (non interrupted utterance), then it is not a DM". The scores are therefore very low: CCI $= 67.5\%, \kappa = 0.09, r = 0.98, p = 0.67, f = 0.80$.

As was the case when INITIAL was replaced by PAUSE-BEFORE, the replacement of FINAL by PAUSE-AFTER slightly increases the scores to CCI $= 75.8\%, \kappa = 0.50, r = 0.74, p = 0.87, f = 0.80$. The resulting tree consists of the following rules: "*well* is not a DM when it ends a completed utterance, and it is a DM in all other cases" and "*like* is a DM only when it is followed by a pause longer than 60 ms, or when it is final in an interrupted utterance". The limit of the pause after *like* corresponds to a value situated between the average values of pauses after DMs and non DMs shown in Table 6.4 above. Together with the similar value of PAUSE-BEFORE, these values match our intuitions about the prosodic markedness of DMs in spoken utterances.

Finally, an improvement is observed when all the three prosodic/temporal features are used, leading to the best scores of the experiments with positional and prosodic/temporal features shown in this section (see Table 8.6): CCI $= 79.4\%, \kappa = 0.55, r = 0.82, p = 0.86, f = 0.84$. Not only do these scores improve over the scores of classifiers based on one or two features, but also over the corresponding scores obtained by decision trees that use only the positional features TOKEN, INITIAL and FINAL.

Duration

Despite the observations on the duration of DMs made in Figures 6.1 and 6.2 above, DURATION does not appear as a relevant feature at first glance. For instance, when only DURATION is available as a feature, the decision tree constructed by C4.5 does not make use of it, and consists of the baseline majority classifier ("all occurrences are DMs") whose scores are given in Table 8.1 above.

When TOKEN is made available in addition to DURATION, the scores improve slightly and the DURATION feature appears in the decision tree, whose scores are: CCI $= 71.9\%, \kappa = 0.37, r = 0.79, p = 0.78, f = 0.79$. In terms of κ, this performance is in fact lower than using TOKEN alone (cf. last line of Table 8.1 with κ=0.72) while the number of CCIs improves slightly. Moreover, when adding DURATION to TOKEN, PAUSE-BEFORE and PAUSE-AFTER, the performance does not improve at all.

To explore even further the impact of DURATION on DM identification, C4.5 can be trained using this feature separately on the occurrences of *like* and those of *well*. The resulting classifiers show the existence of a very low correlation of DURATION with the DM / non DM distinction, along the lines observed in Figures 6.1 and 6.2 above.

Considering first the case of *like*, the best classifier using only DURATION reaches CCI $= 56.9\%, \kappa = 0.11, r = 0.40, p = 0.53, f = 0.46$. The decision tree consists of the following rules: "*like* is a DM if its duration is longer than 220 ms and shorter than 350 ms, and is not a DM

otherwise". This interval is almost identical to the interval inferred from the observations in Figure 6.1 above, which showed that DMs *like* slightly outnumbered non DMs in the [220 ms; 360 ms] duration interval. This correlation is however very week since $\kappa = 0.11$ only.

A similar experiment with *well* indicates that the influence of DURATION is even smaller than in the case of *like*, because the best decision tree found by C4.5 using DURATION remains the majority classifier ("all occurrences are DMs") which does use the feature at all. This is also due to the small proportion of non DMs *well* (497 out of 4,136), which leads to a high score for the majority classifier (CCI = 88.0%, but κ=0).

To conclude on positional and prosodic/temporal features, the feature selection studies in Section 8.4 below show that prosodic/temporal features are less useful than lexical ones, but slightly more than positional features (which are logically subsumed by the prosodic features). However, since human annotators performed significantly better when allowed to use sound files, some prosodic features are probably crucial for the identification of DMs independently of the lexical information. Further work on prosody, in particular on pitch variation, is necessary to determine these features.

8.3.3 Correlation of DM Use with Sociolinguistic Features

If *like* and *well* are not distinguished, the sociolinguistic features alone do not permit the construction of a classifier with a non-zero score. When the two types are distinguished using the TOKEN feature, the best decision tree generated by C4.5 using all sociolinguistic features is the majority classifier for *well* ("all occurrences of *well* are DMs") and a more refined classifier for *like*. Indeed, a number of heavy DM-*like* users are identified; the classifier considers as DMs all occurrences of *like* that they produce, while all occurrences produced by other speakers are considered non DMs. The other sociolinguistic features are not used in this case. This simple classifier has the following performance: CCI = 77.3%, $\kappa = 0.47, r = 0.88, p = 0.80, f = 0.84$. These are well above the baseline scores obtained using TOKEN only, shown in the last line of Table 8.1.

The sociolinguistic features were also explored one by one, but no significant influence of the NATIVE feature was found (i.e. no decision tree using only this feature scores above $\kappa = 0$. For all other features, no correlation was found when *like* and *well* are not distinguished, though some relations appear when each DM is considered separately, a fact that shows that the influence of sociolinguistic features is token-specific.

A number of sociolinguistic features appear to be relevant in the case of *like*, considering only the occurrences of *like*. Using EDUCATION, the best tree found by C4.5 reaches $\kappa = 0.39$, the baseline being here $\kappa = 0$. This tree corresponds to the following rule: "if the speaker is an undergraduate or a graduate, consider all tokens of *like* as DMs; if the speaker is a post-doc or a professor, consider all tokens of *like* as non-DMs".

A similar correlation (classifier reaching $\kappa = 0.40$) holds for the (region of) ORIGIN feature, as the following rule was found: "if the speaker is from the US West, consider all tokens of *like* as DMs; otherwise, consider all tokens of *like* as non DMs". An even stronger correlation ($\kappa = 0.44$) is observed for AGE: "if the speaker is under 30, consider all tokens of *like* as DMs; otherwise, consider them as non DMs". These two experiments bring statistical evidence that younger speakers from the US West tend to overuse *like* as a DM, which corroborates a view commonly held by linguists, who often consider the DM *like* as a feature of adolescent speech—see for instance Andersen (2001). Since in the ICSI-MR data there were a majority of speakers under 30 from the US West, which did not hold a PhD at the time of recording, it is not possible to identify the precise feature that correlates with DM-*like* overuse among AGE, ORIGIN or EDUCATION).

These results illustrate how this approach to automatic identification of DMs can be used to study speaker-biases in DM use: the C4.5 learning method complements the purely quantitative measures described in Section 3.3 above, and will be enriched below with a specific study of the relevance of each feature separately in Section 8.4. The best DM classifiers do not seem to require sociolinguistic features, but such information, when available, could be used to increase the accuracy of DM recognition.

8.3.4 Correlation of DM Use with Dialogue Acts

As indicated above in section 6.4, there are two possible ways to use the dialogue act (DA) information for each utterance containing a DM candidate: either by decomposing each DA label and using the tags as features, or by keeping the DA labels as they occur, with a frequency threshold to avoid too many possible values.

In both cases, experiments with the DA feature alone do not indicate that this feature is useful for DM identification. For instance, when using only the TOKEN and the DA features, the results are not improved over the case when TOKEN is used alone, as a baseline: the number of CCIs slightly increases, but κ slightly decreases. The scores with DA and TOKEN are: CCI = $71.9\%, \kappa = 0.38, r = 0.78, p = 0.79, f = 0.78$, while with TOKEN only the scores are: CCI = $70.5\%, \kappa = 0.42, r = 0.64, p = 0.88, f = 0.74$ (copied from Table 8.1 above).

A series of experiments was also conducted with two classifiers other than decision trees, namely the already mentioned Bayesian Network and a discrete decision tree learner called Id3. The second method allows that some instances cannot be classified by the tree, which is why the values of κ are quite high, while the number of CCIs is lower than before. The main goal of these experiments was to compare the merits of the DA features. The method is similar to the one described at the beginning of Section 8.3 above, and compares separately for *like* and *well* the scores obtained in three conditions:

1. using all the available features;
2. using the DA features only;
3. using all the other features except DA.

The results are summarized in Table 8.7 using the label-based representation of DAs and considering only labels that occur more than 20 times in the data. These results show that the DA features do not have any visible contribution to the best scores, and lead to classifier scoring barely above baseline when used alone.

The results above are only partly surprising. Indeed, even if DMs play a role in discourse structure, it is not clear that they should be related to dialogue acts. For instance, *well* can equally introduce a question, a command, a statement, etc., without being correlated to any of these specific DAs[10].

It is in fact more likely that the two DMs studied here are related to the topic structure or to the negotiation of the information context, which are probably not captured by the DA annotation. In addition, the function of DM *like* could pertain to a more local level than the utterance level captured by DA labels[11]. It is also possible that the DA feature is not used optimally, and that some of the lower-tier DA tags that occur infrequently are in fact highly discriminative for the two DMs considered here.

[10] A study of the lexical correlated as the one conducted here on lexical features could determine whether, among the lexical items mostly correlated with certain DAs, there are DMs or not.

[11] Note however that using DAs leads to $\kappa = 0.12$ for *like*, which is higher than the κ for *well* obtained above.

Token	Method	Features	CCIs (%)	κ
like	BayesNetK2	All	84.7	0.69
		Without DA labels	84.5	0.69
		Only DA labels	57.5	0.12
	Id3	All	60.3	0.69
		Without DA labels	60.5	0.69
		Only DA labels	57.4	0.12
well	BayesNetK2	All	97.4	0.87
		Without DA labels	97.4	0.87
		Only DA labels	88.1	0.01
	Id3	All	88.2	0.87
		Without DA labels	88.3	0.86
		Only DA labels	88.1	0.01

Table 8.7: DM identification scores with and without DA labels on utterances, studied separately for *like* and for *well*. The Id3 decision tree is allowed to leave unclassified instances.

The best overall result for these experiments, with respect to the κ score, is obtained using a Bayesian network classifier and all the available features: κ=0.783±0.016. The score obtained without using the DAs is κ=0.780±0.016, which is not significantly lower than the previous one in terms of the 95% confidence interval.

8.4 Automatic Attribute Selection

Attribute selection algorithms compare the merits of features for DM identification in a more systematic way. These algorithms offer an additional source of information, which can corroborate the conclusions arrived at so far by (1) computing descriptive statistics about the data (Chapter 6) or (2) by looking at the performance of classifiers (Section 8.3).

Two main methods implemented in the Weka toolkit are used in this section. The *correlation-based feature subset selection (CFS)* aims at finding the best subset of features for DM identification by examining the individual predictive power of each feature (with respect to DM vs. non DM identification) while at the same time minimizing redundancy in the subset, i.e. avoiding to keep in the same subset features that are too correlated. The result of this search procedure, which uses a best-first algorithm with backtracking, is an optimal feature subset which has both high predictive value and low redundancy.

A second method provides independent relevance scores for each feature, regardless of their use in a possible subset, which allow an overall ranking of all the features. Two possible relevance criteria are the *information gain* of the feature with respect to the DM / non DM classification (the entropy of the DM class minus the conditional entropy of the DM class given the feature), and the χ^2 statistic of the feature with respect to the DM class (Witten and Frank, 2000). The rankings provided by these methods appeared to be very similar, so only the information gain is used here. To quantify differences between features, numerical values will be also provided with the rankings.

Both attribute selection methods can be applied to various representations of features for DM identification. The main difference that we explored is related to the two possible encodings of lexical features (Section 6.1.3 above). When lexical collocations are encoded by position-related variables (WORD(−1), WORD(+1), etc.), whose possible values are individual words, the number of features to be examined by attribute selection is quite low as there are $2N$ lexical features

for a window of $[-N, +N]$ around the candidate. This first option (Section 8.4.1 below) will be used to compare the importance of lexical, positional, prosodic and sociolinguistic features. The second encoding option uses one variable per word (for words above a certain frequency threshold only), whose value is the position of the word with respect to the candidate DM. This representation is useful to compute the most discriminative collocations and their position (Section 8.4.2 below).

8.4.1 Comparative Relevance of the Features

The CFS algorithm indicates that the following subset of attributes is optimal: {TOKEN, PAUSE-BEFORE, INITIAL, WORD(−1)}. This result confirms quite clearly the previous observations: lexical collocations are a key feature, as is the distinction of the two DMs, and the use of the pause before the candidate, which codes also for its utterance-initial character. In addition, this result indicates that the collocations in which the candidate DM is the second word are the most discriminative ones. It is somewhat surprising to find both INITIAL and PAUSE-BEFORE since the former feature is entirely predictable from the latter. However, it is likely that the search method cannot find this relation, given that PAUSE-BEFORE is a numerical feature—a value of '10 s' indicating that the candidate DM is utterance initial—while INITIAL is discrete ('y' or 'n' values).

The ranking of each feature using information gain or χ^2 yields the same ranking and a similar variation in scores, though the absolute values of the scores are very different: the most discriminative feature is the word before the candidate, WORD(−1), with an information gain of 0.52 and a χ^2 value of 5268. The order of the features and their information gain is represented graphically in Figure 8.2. The values indicate that by far the most informative feature is the word before the candidate DM WORD(−1) (a feature that has many possible values, thus increasing informativeness), followed at some distance by PAUSE-BEFORE, INITIAL, WORD(+1) (the word after the candidate) and TOKEN. These features are all included in subset found by CFS, apart from WORD(+1). The remaining features exhibit much lower correlations with the DM class, ending with GENDER at 0.0005[12].

Similar results appear if CFS attribute selection is applied separately for *like* and for *well*. The best subset for *like* is: PAUSE-BEFORE, PAUSE-AFTER, WORD(−1), WORD(+1), and SPEAKER. However, if these features are used by C4.5 to construct a decision tree, the prosodic/temporal features are mainly used as indicators of utterance-initial and utterance-final (completed utterance) positions. The best subset for *well* is quite similar: PAUSE-BEFORE, INITIAL, FINAL, and WORD(−1). The presence of SPEAKER in the optimal set for *like* indicates that some of the participants have marked preferences for using *like*, while no such phenomenon is detected for *well*.

Conversely, the analyses of features based on information gain or on χ^2 lead to similar lists for *like* and *well* considered separately. For *like*, the three most informative features are WORD(−1), WORD(+1) and SPEAKER, while all the following ones have much lower information gains. For *well*, the three most informative features are WORD(−1), PAUSE-BEFORE and INITIAL, which are followed very closely by WORD(+1), PAUSE-AFTER and FINAL. The full lists of features and their information gain values are gathered in Table 8.8.

The best feature subset found by CFS does not obtain a score that is close to the best scores obtained using the C4.5 learner. The best classifier found by C4.5 using TOKEN, PAUSE-BEFORE, INITIAL, WORD(−1) reaches the following scores: CCI = 86.8%, $\kappa = 0.69, r = 0.96, p =$

[12]The actual (rounded) values are: WORD(−1) 0.5167, PAUSE-BEFORE 0.2601, INITIAL 0.2226, WORD(+1) 0.1947, TOKEN 0.1550, SPEAKER 0.0385, PAUSE-AFTER 0.0358, FINAL 0.0250, DURATION 0.0196, AGE 0.0126, EDUCATION 0.0059, COUNTRY 0.0052, NATIVE 0.0026, GENDER 0.0005.

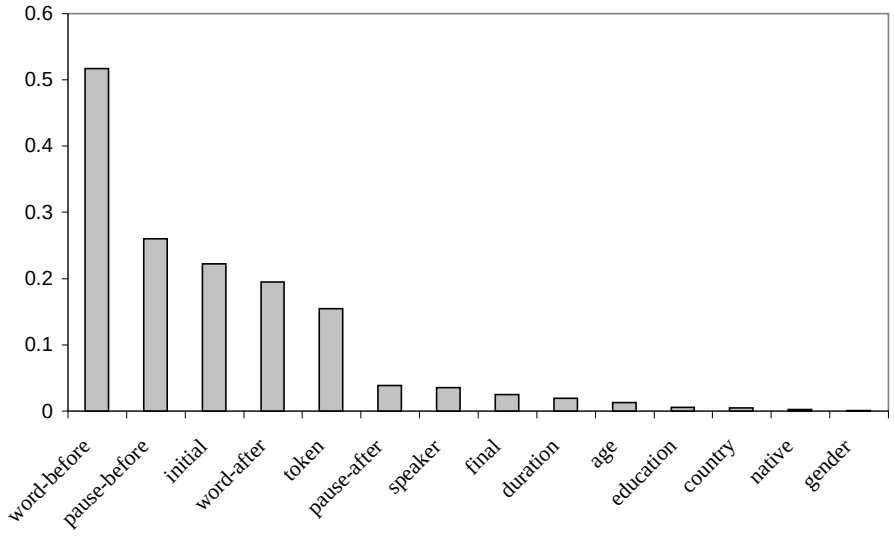

Figure 8.2: Information gain for the features used to identify DMs.

Like		*Well*	
Feature	Information gain	Feature	Information gain
WORD(−1)	0.4406	WORD(−1)	0.3924
WORD(+1)	0.2135	PAUSE-BEFORE	0.2302
SPEAKER	0.1518	INITIAL	0.1956
PAUSE-BEFORE	0.0609	WORD(+1)	0.1503
AGE	0.0604	PAUSE-AFTER	0.1038
PAUSE-AFTER	0.0492	FINAL	0.1037
EDUCATION	0.0404	SPEAKER	0.0426
INITIAL	0.0357	DURATION	0.0332
COUNTRY	0.0174	AGE	0.0097
FINAL	0.0148	COUNTRY	0.0046
GENDER	0.0140	EDUCATION	0.0041
DURATION	0.0122	NATIVE	0.0009
NATIVE	0.0010	GENDER	0.0005
TOKEN	0	TOKEN	0

Table 8.8: Separate ranking of features for *like* and *well* according to their information gain. Significant decreases in scores are indicated for each marker. The TOKEN feature is irrelevant when the two DMs are considered separately.

$0.86, f = 0.91$, so κ is especially lower than for the best classifiers ($\kappa = 0.78$ in Table 8.4). However, when using a larger set of features obtained by merging the best subsets for *like* and for *well*, the scores increase significantly to reach values that are much closer to the best ones: CCI $= 88.4\%, \kappa = 0.73, r = 0.97, p = 0.87, f = 0.92$. The fact that the best subset does not allow a decision tree to reach the best possible scores is probably due to the difference between the two methods: the search for the best subset looks only at correlation between features and the DM / non DM classification, which amounts to a much simpler classifier than the C4.5 decision trees or the Bayesian network.

8.4.2 Most Relevant Lexical Indicators of DMs

Attribute selection can be used to determine the most discriminative collocations, i.e. the words that appear around a candidate DM and indicate that the candidate is likely to be either a DM or a non DM[13]. To be able to assess the discriminative power of individual words, these must be represented as features, whose possible values are the positions of the respective words relative to the candidate.

A number of lexical indicators appear in the best feature sets found by CFS, depending on the candidate DM for which the search is done:

- for *like* and *well* together, five lexical items appear along with TOKEN, PAUSE-BEFORE, INITIAL. Their roles in DM identification are made explicit in parenthesis: *something* (precedes non DM *like*), *things* (same function), *seems* (same function), *that* (follows non DM *like*), *to* (follows non DM *like* and precedes DM *like*). The five lexical items found by CFS thus concern only the identification of DM *like*.

- for *like* alone, the two most relevant features are PAUSE-BEFORE and PAUSE-AFTER, together with the five lexical items just mentioned.

- for *well* alone, the subset of most relevant features is: PAUSE-BEFORE, INITIAL, FINAL, plus two lexical indicators, *as* and *very* (both precede non DM occurrences of *well*).

The individual ranking of the lexical indicators is easier to interpret when ranked separately for *like* vs. *well*, as in Table 8.9. The table shows the upper part of the rankings, including all non-lexical features except NATIVE, and of course TOKEN, which is irrelevant when the ranking is done separately for each DM type.

The scores reached using only the most informative lexical indicators are of course lower than the best scores, or the ones reached in the previous subsection. Indeed, when the WORD(-1) feature is used, the available information concerns all possible collocations before the candidate, which is not the case with in the present approach, for which the more individual collocations are available, the higher the classification performance. Overall, the lexical features appear to have considerable importance, especially for *like*—which is also, in our data, the most ambiguous of the two DMs—along with prosodic/positional ones. Some sociolinguistic features also rank quite high, especially for *like*: for instance, knowledge of the speaker is the most important feature for the identification of DM *like* in the absence of WORD(-1) and WORD$(+1)$, which are replaced in this experiment by a multitude of lexical features, which all have less importance than the SPEAKER. The experiment in the previous subsection (Table 8.8) shows nevertheless that the cumulated information gain of the lexical information is considerably higher than the gain of the SPEAKER feature. *Like* appears nevertheless to be much more prone to individual or sociolinguistic biases than *well*.

[13] A number of *a priori* observations and a list of words appeared in Section 6.1 above.

Like		*Well*	
Feature	Information gain	Feature	Information gain
SPEAKER	0.152	PAUSE-BEFORE	0.230
that	0.093	INITIAL	0.196
something	0.075	PAUSE-AFTER	0.104
PAUSE-BEFORE	0.061	FINAL	0.104
AGE	0.060	*as*	0.050
PAUSE-AFTER	0.049	SPEAKER	0.043
to	0.046	DURATION	0.033
EDUCATION	0.040	*very*	0.027
things	0.036	*how*	0.021
INITIAL	0.036	*be*	0.011
seems	0.033	*pretty*	0.010
would	0.022	AGE	0.010
COUNTRY	0.017	*on*	0.009
this	0.017	*may*	0.007
have	0.016	*could*	0.007
FINAL	0.015	*for*	0.006
was	0.014	*work*	0.006
looks	0.014	*with*	0.005
GENDER	0.014	COUNTRY	0.005
of	0.014	*do*	0.004
DURATION	0.012	EDUCATION	0.004
sounds	0.011	*really*	0.003

Table 8.9: Ranking of lexical and non-lexical features for *like* and *well* according to their information gain. Only the initial part of the list is displayed here, including all non-lexical features except for NATIVE and GENDER.

8.5 Discussion

The results of the experiments with DM identification presented here can be interpreted from a number of perspectives. First, the best scores must necessarily be compared with baseline ones and with inter-annotator agreement values in order to assess the overall accuracy of automatic identification (Section 8.5.1). Then, the results can be used to shed light on the relative importance of the features used for identification, and in particular on the necessity to process the two DMs studied here differently, using the TOKEN feature (Section 8.5.2). Finally, the presents results will be shown to compare favourably with previous results ones for DM identification (Section 8.5.3).

8.5.1 Overall Assessment of Scores

The best classification scores obtained in these experiments for both *like* and *well* are approximately: CCI $= 90\%, \kappa = 0.78, r = 0.96, p = 0.90, f = 0.93$. These are average values on the ten folds used in the cross-validation procedure, which appears to be particularly stable, as the 95% confidence intervals for the previous values are $\pm 2\%$ for κ and less than $\pm 1\%$ for the other scores. The best results are obtained—with overlapping 95% confidence intervals—by a Bayesian network, a decision tree trained with C4.5, and a support vector machine.

The best scores are well above the baseline ones, although depending on how the baseline is defined, some very simple classifiers have scores that are well above zero. For instance, the scores of a majority classifier ('DM') are approximately: CCI $= 66\%, \kappa = 0, r = 1, p = 0.66, f = 0.79$, and the scores of a token-specific majority classifier ("*like* is never a DM, *well* is always a DM") are: CCI $= 71\%, \kappa = 0.42, r = 0.64, p = 0.88, f = 0.74$. However, a more informed baseline which uses the TOKEN and WORD(-1) features already reaches CCI $= 87\%, \kappa = 0.72, r = 0.91, p = 0.90, f = 0.90$. The range of variation of scores that are significantly above baseline is much smaller than the [0; 1] interval.

The best identification scores are quite dissimilar for *like* and for *well*, and baseline scores are here much lower too. For *like*, the best scores are around CCI $= 85\%, \kappa = 0.69, r = 0.90, p = 0.79, f = 0.84$ while the baseline (majority classifier) is CCI $= 55\%, \kappa = 0, r = 0, p = 0, f = 0$—recall and precision are both zero as they are calculated for the DM identification task, and the majority class for *like* is 'non DM'. For *well*, the best scores are much higher, at around CCI $= 98\%, \kappa = 0.88, r = 0.99, p = 0.98, f = 0.99$ while the baseline (majority classifier) is also much higher (except for κ) at: CCI $= 88\%, \kappa = 0, r = 1, p = 0.88, f = 0.94$.

The best scores obtained are in fact comparable to inter-annotator agreement values obtained in Chapter 4, which reached $\kappa = 0.74$ for the best experimental conditions, when subjects used transcripts and audio recordings of the dialogues. This similarity indicates that automatic classifiers have probably reached the highest possible performance in the present experiments, and therefore that the set of features that was used was sufficient to reach an accuracy comparable to human annotators. Improving the scores seems thus to require also a more reliable annotation, obtained for instance by allowing experienced annotators to discuss and adjudicate their initial annotations.

8.5.2 Relevance of Features to DM Identification

Experiments with separate subsets of features, as well as attribute selection algorithms, allow us to posit a clear ranking of the utility of the various features.

The most important features appear to be the *lexical* ones: lexical collocations that can be learned from the training data are the most reliable indicators of DMs (or non DMs). Among these, it appeared that the word before a candidate DM (WORD(-1) feature) is the most useful

one, especially as it implicitely codes the utterance-initial character as well. Scores obtained using lexical features exclusively are less than 5% close to the best overall scores. Decision trees based only on lexical features (or even on TOKEN and WORD(−1) only) are nearly optimal. It is therefore surprising that these features were not used in Litman's 1996 study. One might hypothesize that the size of Litman's data did not allow these features to be learned reliably. In our first study Zufferey and Popescu-Belis (2004), lexical collocation rules were defined by the experimenters, and only in the present ones are they learned automatically—a method that appears to be more efficient than manual definition.

Positional and prosodic/temporal features are significantly less efficient than lexical ones, when used alone, although they appear in the best decision trees just below lexical features. The utterance-initial, and then the utterance-final character of the DM candidate appear to be correlated with the DM / non-DM distinction, provided the tokens are classified separately. Even more reliable is the use of PAUSE-BEFORE and PAUSE-AFTER, and especially their joint use. Machine learning determines that pauses of ca. 60 ms around a token tend to characterize DMs, whereas the duration of the token itself is almost irrelevant to its classification.

The *sociolinguistic features* are only slightly correlated to DM use, almost exclusively for DM *like*: the most reliable indicators are the identity of the speaker and their education level. However, these features are much less reliable for DM detection than the other ones, which supersede them almost entirely. As for *dialogue acts*, no reliable correlation was observed in this study, for reasons that were hypothesized and discussed above.

Finally, the present experiments show that the TOKEN feature is crucial to DM identification, shows that the two DMs studied here, *like* and *well* are much better classified individually than as a unique class. It appears therefore that it is better to process DM candidates separately according to their type, a conclusion that matches the theoretical insights showing that DMs are not an homogeneous class ("connectives" aren't one either). Although some of the previous features do generalize to both tokens (such as PAUSE-BEFORE), many of the most relevant features are token-specific, in particular lexical features.

However, if the learning algorithms are allowed to use TOKEN, and if enough training examples are available for each ambiguous token, then it is not particularly difficult to learn token-specific features from the data and apply them to DM identification. There is thus no particular need to look for general-purpose features, and not many chances that such features can indeed be found for the several dozen types of ambiguous DM candidates.

This conclusion echoes the summary provided by Passonneau and Litman (1997) of Litman's study:

> "Initial phrase position [...] was correlated with discourse signaling uses of cue words in (Hirschberg and Litman, 1993). [...] Litman (1996) found that treating cue phrases individually rather than as a class enhanced the results of (Hirschberg and Litman, 1993)."

8.5.3 Comparison with previous work

Previous work on DM identification was outlined in Chapter 2. Table 8.10 summarizes not only the DM identification scores that were reached by previous experiments focussed on DM identification, but also the other parameters that are important for comparison purposes: the number of DM types and tokens that were used for training (if applicable) and test, the number of human annotators, and the method and features. Conversely, Table 8.11 provides the corresponding characteristics and scores of our experiments, starting with (Zufferey and Popescu-Belis, 2004) and continuing with the overall and token-specific scores of the present experiments.

Study	Annotators	DM types	Candidates	Method and features	Results
HL93	1	*now*	100	intonational hand-coded	C=98%
"	"	*now*	100	punctuation	C=82%
"	"	*well*	52	intonational hand-coded	C=98.1%
"	2 (8% tokens discarded)	34	878	intonational hand-coded	C=75.4% R=63.1% P=88.3% κ=0.52
"	"	34	878	punctuation	C=80.1% R=57.3% P=82.6% κ=0.54
"	non conjuncts	31	495	intonational, hand-coded	C=85.3% R=82.7% P=81.5% κ=0.69
SM94	1	34	1,027	baseline decision tree, punctuation	C=79.16%
"	"	34	1,027	decision tree built by GA, punctuation, lexical features	C=79.20% (58-fold)
L96	same as HL93	34	878	CGRENDEL or C4.5 classifiers using prosodic and textual features	C=83.1%±3.4% (10-fold)
"	non conjuncts	34	495	"	C=83.4%±4.1%
"	same as HL93	34	878	prosodic phrase	C=85.5%±3.4%
"	non conjuncts	34	495	"	C=87.4%±3.3%
"	same as HL93 including ambiguous tokens	34	953	prosodic, textual	C=77.6%±4.1%
HA99	1 (DMs only)	$\sim$ 30	8,278 DMs	language model with POS	for DMs only C*=93.57% R=97.26% P=96.32%

Table 8.10: Synthesis of DM identification methods and results. The studies are abbreviated as follows: HL93 for (Hirschberg and Litman, 1993); SM94 for (Siegel and McKeown, 1994); L96 for (Litman, 1996); and HA99 for (Heeman and Allen, 1999). The evaluation metrics are abbreviated as follows: C for accuracy or correctly classified instances (inverse of error rate), R for recall and P for precision. Double quotes indicate the same value as in the above cell.

Study	Annotators	DM types	Candidates	Method and features	Results (10-fold c.v.)
ZPB04	1–2	2	8,655	C4.5 decision trees, hand-crafted collocation rules	C=89.2% R=98.0% P=87.1% κ=0.75
PBZ	1–2	2	8,655	C4.5 decision trees, lexical, temporal and sociolinguistic features	C=90.5%±0.6% R=95.7%±0.4% P=90.4%±0.8% κ=0.78±0.02
”	”	*like*	4,519	”	C=84.5%±1.0% R=89.9%±1.5% P=78.9%±0.8% κ=0.69±0.02
”	”	*well*	4,136	”	C=97.5%±0.5% R=99.3%±0.3% P=97.9%±0.4% κ=0.88±0.02

Table 8.11: Results of the present study (PBZ), in comparison with our previous study (ZPB04, Zufferey and Popescu-Belis, 2004). The evaluation metrics are abbreviated as follows: C for accuracy or correctly classified instances (inverse of error rate), R for recall and P for precision.

The present results compare favourably with previous ones obtained in similar conditions. As noted earlier, the accuracy or number of correctly classified instances is not always an eloquent indicator of performance, as the baseline accuracies are already very high. We prefer here the use of κ, and our κ scores are significantly higher than those obtained in other studies. Overally accuracy is also influenced by the DM types considered in the study, as shown for instances by the differences in scores for *like* vs. *well*. Therefore, comparison is not always significant if the studies do not consider the same DM candidates. Of course, the availability of a large annotated data set would allow more significant comparison of DM identification algorithms, but such a resource, annotated for dozens of DMs, does not yet exist. As indicated above, the present data, though limited to two highly ambiguous DMs, is available for other studies.

Chapter 9

Conclusion and Perspectives

This work has presented a full panorama of DM disambiguation. It has introduced the notion, explained why it is useful to computational linguistics, discourse modelling and NLP, and outlined the previous attempts to recognize DMs. The annotation of DMs was then discussed, followed by the development of a classifier for DMs like and well. This classifier improves performance over previous studies, and allows the identification of a number of useful features, among which token-specific lexical features appear to be the most important ones.

9.1 Main Results and Lessons Learned

This article discussed the disambiguation of DM-candidates, focusing on two highly ambiguous tokens: *like* and *well*. We first outlined their ambiguity and their role as DMs, as well as their relevance to other NLP tasks. Human identification of DMs is quite consistent, provided audio information is available; in this case, inter-annotator agreement reaches $\kappa = 0.74$. A general-purpose POS tagger scores well below this level. Therefore, we used C4.5 learning of decision trees on data annotated with ca. 8,500 occurrences of *like* and *well*, and obtained classifiers reaching $\kappa = 0.78$, and 90% accuracy. The most efficient features are collocation-based filters, constructed from co-occurrence studies and linguistic insights. The automatic construction of decision trees also shows indirectly the influence of speaker-related parameters such as age and place of origin on DM use.

9.2 Future Work

Future work should focus first on the generalization of the features to other ambiguous DM-candidates, such as *you know*. This requires manual annotation of a sizeable amount of instances for training and test, and adaptation of the features, a cost which should be assessed. More elaborate prosodic features should be added, provided they can be detected automatically for each DM-candidate using the sound files. Although POS tagging seems quite unreliable on DMs, the results of a tagger trained on speech could be used at least to generalize collocation-based features.

Given the strong pragmatic function of DMs, it is unlikely that low-level features combined with machine learning will entirely solve the problem. It may seem that the pragmatic role of a DM could be confirmed only by a full semantic analysis of an utterance, which is still a remote goal for NLP in domain-independent cases. However, a full semantic analysis of an utterance seems to require in its turn the disambiguation of DM-candidates. It appears thus that DM disambiguation must be done more or less in parallel to semantic analysis, either within an

entirely new analysis paradigm, or, more likely, by bootstrapping the two processes with low-level features, then propagating information from one to another. The technique described in this article offers such a set of low-level features for DM disambiguation.

Acknowledgments

The research presented here was supported by the Swiss National Science Foundation through its NCCR on Interactive Multimodal Information Management, or (IM)2 (see `http://www.im2.ch`). The investigation of DMs is part of the Shallow Dialogue Analysis (SDA) model developed first in the (IM)2 sub-project on Multimodal Dialogue Management and now in the sub-project on Multimodal Content Abstraction (see `http://www.issco.unige.ch/projects/im2/`). We would like to thank Liz Shriberg and Barbara Peskin from ICSI for their help with the ICSI Meeting Corpus and its annotations, as well as several participants to the SIGDial 2004 workshop at MIT for their comments on preliminary results. We also thank Dan Jurafsky and Oliver Mason for answering our questions about, respectively, the Switchboard data and the QTag POS tagger.

Bibliography

Karin Aijmeer, 2002. English Discourse Particles: Evidence from a Corpus. John Benjamins, Amsterdam.

Gisle Andersen, 2000. The role of the pragmatic marker 'like' in utterance interpretation. In: Gisle Andersen, Thorstein Fretheim (Eds.), Pragmatic Markers and Propositional Attitude. Pragmatics and Beyond New Series 79. John Benjamins, Amsterdam, pp. 17–38.

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

Diane Blakemore, 2002. Relevance and Linguistic Meaning: the Semantics and Pragmatics of Discourse Markers. Cambridge University Press, Cambridge, UK.

Gillian Brown, George Yule, 1983. Discourse Analysis. Cambridge University Press, Cambridge, UK.

Donna K. Byron, Peter A. Heeman, 1997. Discourse marker use in task-oriented spoken dialogues. In: Eurospeech '97. Rhodes, Greece, pp. 2223–2226.

Jean Carletta, 1996. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22 (2), 249–254.

Jacob Cohen, 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Robin Cohen, 1984. A computational theory of the function of clue words in argument understanding. In: Coling-ACL 1984 (10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics). Stanford, CA, USA, pp. 251–258.

Marie-Hélène Corréard, Valerie Grundy (Eds.), 1994. The Oxford-Hachette French Dictionary: French-English English-French. Oxford University Press, Oxford, UK.

Richard Craggs, Mary McGee Wood, 2005. Evaluating discourse and dialogue coding schemes. Computational Linguistics 31 (3), 289–295.

Vivian de Klerk, 2005. Procedural meanings of *well* in a corpus of Xhosa English. Journal of Pragmatics 37, 1183–1205.

Barbara Di Eugenio, Michael Glass, 2004. The kappa statistic: A second look. Computational Linguistics 30 (1), 95–101.

Alain Duval, Lorna Sinclair Knight (Eds.), 1995. Collins-Robert French-English English-French Dictionary, Unabridged. HarperCollins Publishers, Glasgow, UK, 4th edition.

Bruce Fraser, 1990. An approach to discourse markers. Journal of Pragmatics 14, 383–395.

Bruce Fraser, 1996a. Pragmatic markers. Pragmatics 6 (2), 167–190.

Bruce Fraser, 1996b. What are discourse markers? Journal of Pragmatics 31, 931–952.

Janet M. Fuller, 2003. The influence of speaker roles on discourse marker use. Journal of Pragmatics 35 (1), 23–45.

Barbara J. Grosz, Julia Hirschberg, 1992. Some intonational characteristics of discourse structure. In: ICSLP 1992 (2nd International Conference on Spoken Language Processing). Banff, Canada, pp. 429–432.

Barbara J. Grosz, Candace L. Sidner, 1986. Attentions, intentions and the structure of discourse. Computational Linguistics 12 (3), 175–204.

Peter A. Heeman, 1997. Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog. Ph.D. thesis, University of Rochester, Rochester, NY.

Peter A. Heeman, James F. Allen, 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers utterances in spoken dialogue. Computational Linguistics 25 (4), 1–45.

Peter A. Heeman, Donna Byron, James F. Allen, 1998. Identifying discourse markers in spoken dialog. In: Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing. Stanford, CA.

Julia Hirschberg, Diane Litman, 1993. Empirical studies on the disambiguation of cue phrases. Computational Linguistics 19 (3), 501–530.

Julia Hirschberg, Christine H. Nakatani, 1996. A prosodic analysis of discourse segments in direction-giving monologues. In: ACL 1995 (34th Annual Meeting of the Association for Computational Linguistics). Santa Cruz, CA, pp. 286–293.

Paul J. Hopper, Elizabeth C. Traugott, 2003. Grammaticalization. Cambridge University Press, Cambridge, UK.

Eduard H. Hovy, 1995. The multifunctionality of discourse markers. In: Workshop on Discourse Markers. Egmond-aan-Zee, The Netherlands, p. 13.

Ben Hutchinson, 2004a. Acquiring the meaning of discourse markers. In: Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics). Barcelona, Spain, pp. 685–692.

Ben Hutchinson, 2004b. Mining the web for discourse markers. In: Proceedings of LREC 2004 (4th International Conference on Language Resources and Evaluation). Vol. II. Lisbon, Portugal, pp. 407–410.

Adam Janin, Don Baron, Jane A. Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, 2003. The ICSI Meeting Corpus. In: Proceedings of ICASSP 2003 (IEEE International Conference on Acoustics, Speech, and Signal Processing). Hong Kong, China.

Andreas Jucker, 1993. The discourse marker 'well': a relevance-theoretical account. Journal of Pragmatics 19, 435–452.

Daniel Jurafsky, James H. Martin, 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Upper Saddle River, NJ.

Alistair Knott, 1996. A data-driven methodology for motivating a set of coherence relations. Ph.D. thesis, University of Edinburgh, Department of Artificial Intelligence.

Klaus Krippendorff, 1980. Content Analysis: An Introduction to Its Methodology. Sage Publications, Beverly Hills, CA.

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

Diane Litman, 1996. Cue phrase classification using machine learning. Journal of Artificial Intelligence Research 5, 53–94.

Diane J. Litman, 1994. Classifying cue phrases in text and speech using machine learning. In: AAAI 1994 (12th National Conference on Artificial Intelligence). Vol. 1. AAAI Press, Seattle, WA, pp. 806–813.

William Mann, Sandra Thompson, 1988. Rhetorical structure theory: toward a functional theory of text organisation. Text 8 (3), 243–281.

Christopher D. Manning, Hinrich Schütze, 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA.

Daniel Marcu, 2000. The Theory and Practice of Discourse Parsing and Summarization. The MIT Press, Cambridge, MA.

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

Marie Meteer, 1995. Dysfluency annotation stylebook for the Switchboard corpus. Working paper, Linguistic Data Consortium.
URL `http://www.ldc.upenn.edu/Catalog/CatalogList/LDC99T42/DFLGUIDE.PS`

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, Bonnie Webber, 2004. The Penn Discourse Treebank. In: Proceedings of LREC 2004 (4th International Conference on Language Resources and Evaluation). Vol. VI. Lisbon, Portugal, pp. 2237–2240.

Nelson Morgan, Don Baron, Sonali Bhagat, Hannah Carvey, Rajdip Dhillon, Jane A. Edwards, David Gelbart, Adam Janin, Ashley Krupski, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, 2003. Meetings about meetings: research at icsi on speech in multiparty conversations. In: Proceedings of ICASSP 2003 (IEEE International Conference on Acoustics, Speech, and Signal Processing). Hong Kong, China.

Rebecca J. Passonneau, Diane J. Litman, 1997. Discourse segmentation by human and automated means. Computational Linguistics 23 (1), 103–140.

Janet B. Pierrehumbert, 1980. The phonology and phonetics of english intonation. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, distributed by the Indiana University Linguistics Club Publications, Bloomington, IN.

Andrei Popescu-Belis, 2005. Dialogue acts: One or more dimensions? Working Paper 62, ISSCO.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, Bonnie Webber, 2004. Annotation and data mining of the Penn Discourse Treebank. In: Proceedings of the ACL 2004 Workshop on Discourse Annotation. Barcelona, Spain, pp. 88–95.

John R. Quinlan, 1993. C4.5: Programs for Machine Learning. Morgan Kaufman, San Francisco, CA, USA.

Rachel Reichman, 1985. Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model). Bradford Books / The MIT Press, Cambridge, MA.

Suzanne Romaine, Deborah Lange, 1991. The use of 'like' as a marker of reported speech and thought: A case of grammaticalization in process. American Speech 66 (3), 227–79.

Ken Samuel, 1999. Discourse learning: An investigation of dialogue act tagging using transformation-based learning. Ph.D. thesis, University of Delaware, Department of Computer and Information Sciences.

Ken Samuel, Sandra Carberry, K. Vijay-Shanker, 1999. Automatically selecting useful phrases for dialogue act tagging. In: Proceedings of PACLING 1999 (4th Pacific Association of Computational Linguistics Conference). Waterloo, Canada.

Deborah Schiffrin, 1987. Discourse Markers. Cambridge University Press, Cambridge, UK.

Lawrence C. Schourup, 1985. Common Discourse Particles in English Conversations: 'Like', 'Well', 'y'know'". Garland, New York and London.

Lawrence C. Schourup, 2001. Rethinking 'well'. Journal of Pragmatics 33, 1025–1060.

Elizabeth Shriberg, Rebecca Bates, Paul Taylor, Andreas Stolcke, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, Carol Van Ess-Dykema, 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech (Special Issue on Prosody and Conversation) 41 (3-4), 439–487.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, Hannah Carvey, 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In: Proceedings of SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue). Cambridge, MA, pp. 97–100.

Eric V. Siegel, Kathleen R. McKeown, 1994. Emergent linguistic rules from inducing decision trees: Disambiguating discourse clue words. In: Proceedings of AAAI 1994 (12th National Conference on Artificial Intelligence). Vol. 1. AAAI Press, Seattle, WA, pp. 820–826.

Muffy E. A. Siegel, 2002. Like: The discourse particle and semantics. Journal of Semantics 19 (1), 35–71.

John Sinclair (Ed.), 2001. Collins COBUILD English Dictionary, 3rd Edition. HarperCollins Publishers, Glasgow, UK.

Dan Sperber, Deirdre Wilson, 1986/1995. Relevance: Communication and Cognition. Blackwell, Oxford, UK.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, Marie Meteer, 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26 (3), 339–371.

Andreas Stolcke, Elizabeth Shriberg, 1996. Automatic linguistic segmentation of conversational speech. In: Proceedings of ICSLP 1996 (4th International Conference on Speech and Language Processing). Philadelphia, PA, pp. 1005–1008.

Maria Teresa Taboada, 2006. Discourse markers as signals (or not) of rhetorical relations. Journal of Pragmatics 38 (4), 567–592.

Elizabeth C. Traugott, 1995. The role of the development of discourse markers in a theory of grammaticalization. In: XIIth International Conference on Historical Linguistics. Manchester, UK.
URL `http://www.stanford.edu/~traugott/papers/discourse.pdf`

Cornelis J. VanRijsbergen, 1979. Information Retrieval. Butterworth, London.

Iain Witten, Eibe Frank, 2000. Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco, CA.

Sandrine Zufferey, 2004. Une analyse des connecteurs pragmatiques fondée sur la théorie de la pertinence et son application au TALN. Cahiers de Linguistique Française 25, 257–272.

Sandrine Zufferey, Andrei Popescu-Belis, 2004. Towards automatic identification of discourse markers in dialogs: The case of like. In: Proceedings of SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue). Cambridge, MA, pp. 63–71.