Token-based Chunking of Turn-internal Dialogue Act Sequences

Piroska Lendvai and **Jeroen Geertzen**

Dept. of Communication & Information Sciences,
Tilburg University, The Netherlands,
{p.lendvai,j.geertzen}@uvt.nl

Abstract

In this study we compare two sequence learning approaches to chunk dialogue acts within a speaker's turn. We assign a dialogue act label to each token in the transcribed speech stream of a dialogue participant, additionally classifying if the token is at the *beginning* of, *inside*, or *outside* that specific dialogue act. Experimental findings show that both our approaches – conditional random fields and memory-based tagging – largely improve over local classification methods, obtaining comparable scores on distinct datasets. We discuss the interplay between transcription granularity of turns and dialogue act chunking.

1 Introduction

Previous supervised learning approaches to dialogue act tagging are typically applied to dialogue units that are pre-segmented on e.g. turn level, utterance level, or functional unit level, the exceptions being (Warnke et al., 1997) and (Zimmermann et al., 2005). However, automatic segmentation into dialogue units is a significant challenge in itself. Even a short speaker turn can contain more than one dialogue act units, for example an agreement in reply to a proposal and an immediate question ('Fine. Which airport?'); on the other hand, multiple turns of the same speaker may feature one single dialogue act, for instance a sequence of statements.

An important aspect of corpus-based approaches is that training data are mostly derived from transcribed speech, where it is common practice to structure the dialogue participants' token stream (typically containing words, but also disfluent elements, non-speech events, symbols for overlapping speech, etc.) into syntactically or semantically complete units, which are then further segmented into turns along speaker change and time line.

In some circumstances of interaction however, like in situations in which interlocutors are under time pressure to communicate, are under stress, or are engaged in a heated discussion, spoken dialogue does not fully proceed in sequence, but often contains simultaneously occurring events, since speakers may cross-react on each other's (incomplete) utterances in a dynamic way. Transcriptions inevitably commit to one or another granularity criterion, and as such superimpose knowledge-based considerations on how to structure dialogue to some extent. In (Traum and Heeman, 1996) the issue of defining utterance units in spoken dialogue is treated extensively.

Arguably, it is easier to automatically assign a dialogue act (DA) to (semantically) complete units than to incomplete ones, and thus the question arises to what extent DA classification generalises across material created by annotation schemes of different DA unit granularity. In the current study we attempt to make the first explorations of this issue by pursuing a boundary-knowledge-lean approach to two differently transcribed dialogue corpora, focusing on turn-internal DA transitions.

The method we advocate is the application of state-of-the art sequence learning approaches to token-based classification of DAs. Our approach

is to perform sequential tagging based on transcribed words and disfluent elements (henceforth: *tokens*) in streams of utterances up to the point of a speaker change (aka turns). Two supervised classifiers, a memory-based tagger and conditional random fields, are trained to identify each element of the word stream as one of a set of DA types, and also whether the token is an initial or an internal element of a larger DA chunk. This approach can be likened to syntactic phrase chunking, and has been shown to work well for identifying disfluent chunks in spontaneous spoken Dutch discourse (Lendvai et al., 2003).

In the following section we describe the corpora employed in this study, and how token sequences and their contextual attributes were derived from transcribed material. Next, the classification procedure is discussed, where we elaborate on sequence learning as a tagging approach, as well as on the measures of evaluating a chunking task. In Section 4 we present our experimental results, followed by a summary of our findings and pointers to future work.

2 Data

The experiments reported in this paper are carried out on two English language datasets drawn from two corpora, each coding dialogue units in a different way: the Monroe corpus and the MRDA corpus.

The *Monroe corpus* (Stent, 2000) consists of human-human, mixed-initiative, task-oriented dialogues about disaster handling tasks. In each dialogue, the interlocutors are engaged in a collaborative problem-solving, mixed initiative interaction, which involved a scenario at an emergency control centre: an instructor (U) receiving incoming information about a disaster, and a remote subject (S) initially knowing nothing about the task. A typical fragment of these interactions is given in Table 1.

For eight dialogues speech has been manually transcribed, segmented into utterances and turns, and annotated with the DAMSL tag set[1], resulting in a data set of 2,897 speaker utterances that are segmented into 1,701 turns (on average 189 turns per dialogue). Each utterance can have multiple communicative functions in four layers (Allen and Core, 1997); there is almost always at least one function assigned to an utterance. The Monroe corpus is annotated with 13 main DA types that can further contain arguments. We worked with the nine labels contained in the forward-looking and the backward-looking dimension of the annotation. These are: `statement, influence-on-listener, influence-on-speaker, info-request, conventional, other, agreement, understanding, answer`.

$S1$	there are [SIL] three people on a stretcher at the airport
$U1$	mm hm
$S2$	then there's one stretcher [SIL] patient [SIL] at [SIL] the mall
$U2$	+ uh huh [SIL] +
$S3$	+ [SIL] and +
$U3$	here was the heart attack right
$S4$	yeah yeah yeah
$S5$	we should get them to the nearest hospital asap

Table 1: An excerpt from the Monroe corpus.

Because of the nature of the DAMSL scheme, the transcribed utterances in this dataset tend to be long, as DA units are segmented in a rather coarse-grained fashion. It can be guessed however, that interaction between the participants is of a more segmented nature, since overlapping speech is marked by numerous turn-internal $+$ symbols in the transcriptions.

The MRDA corpus (Shriberg et al., 2004) is a companion set of segmentations and annotations on the ICSI Meeting Corpus, which consists of 75 non-scenario based meetings that each are roughly an hour in length. On average, there are about six English speakers, native and non-native, per meeting. Most of the meetings were group discussions about the ICSI meeting recorder project itself or on topics on natural language processing. The sample in Table 2 illustrates an interaction with three dialogue participants.

The utterances in the MRDA corpus have been annotated with a modified version of the SWBD-DAMSL

[1] Annotations are publicly available at `http://www.cs.rochester.edu/research/cisd/resources/monroe/`.

$c1$	um ... so far I have thought of it as sort of adding it onto the modeler knowledge module
$c0$	that is the d-
$c3$	hmm
$c0$	ok
$c0$	yeah

Table 2: An excerpt from the MRDA corpus.

tagset (Jurafsky et al., 1997), in which a dialogue act is a combination of at least one general tag, with a variable number of possible specific tags attached. There are 11 general tags. The MRDA corpus has been used in various segmentation and dialogue act classification studies, e.g. (Zimmermann et al., 2005), and as in most of these studies we worked with dialogue act labels grouped into five types: `backchannels (B)`, `disruptions (D)`, `floorgrabbers (F)`, `questions (Q)`, and `statements (S)`, as well as two miscellaneous labels, `(X)` and `(Z)`.

In this corpus tokens from a speaker are segmented into minimal units that are semantically complete, so that a unit always has only one general DA tag assigned to it. Tags in this dataset are thus mutually exclusive, which is a major difference from the Monroe material. The MRDA data contains 51,452 turns (on average 826 turns per dialogue).

It is important to see that due to these fine-grained DA chunks, the speech stream of one speaker tends be transcribed in a much more scattered way along the course of the interaction than in the Monroe corpus. All three utterances from the speaker on channel 0 in Table 2 would have been transcribed as one utterance in the Monroe corpus, because the DAMSL annotation scheme applied there allows for assigning DA labels on different dimensions, so that a statement and a backchannel could be segmented into one unit. But in the MRDA transcriptions, these token streams are considered as separate units, even with a DA unit of a different speaker inserted between them.

There is an abundance of self-interruptions an other type of disfluencies, overlapping speech, and turn-internal silence in both corpora. The latter two elements are also encoded in markedly different ways in the two datasets: the Monroe transcriptions contain these directly as symbols ($+$ and `[SIL]`, respectively) in the token stream, whereas the MRDA material breaks up the token stream along overlapping speech into separate segments, and encodes silence between tokens by time stamps.

There are a number of other differences between our datasets. First, the DA sets in the two corpora overlap to only a small extent, both in their amount and their aspects: `statement` is a DA in both of them, and there is a `Question` DA type in the MRDA and `Information request` in the Monroe corpus, but the mapping between `Backchannel` in MRDA and `Agreement` as well as `Understanding` in the Monroe material is only partial, whereas the other DA types are difficult to relate across corpora. Additionally, the amount of data in the two datasets differs as well: the Monroe dataset is rather sparse, whereas the MRDA corpus provides thousands of examples to the learners. Finally, the Monroe corpus is a two-party interaction with 'giver and follower' type of roles, whereas the MRDA discussions involve many speakers and a more intertwined flow of interaction.

3 A chunking approach to segmenting dialogue acts

3.1 Classifiers

For the joint learning of the segmentation and labeling, we used two different sequence-based machine learning techniques: *conditional random fields (CRFs)* and *memory-based tagging (MBT)*. Both of these have been shown to be particularly suitable for sequential natural language processing tasks such as part-of-speech (POS) tagging.

CRFs (Lafferty et al., 2001) are probabilistic learners for labeling and segmenting structured data. The algorithm defines a conditional probability distribution over label sequences given a particular observation sequence (in our case a sequence of tokens), rather than a joint distribution over both label and observation sequences. The main advantage of CRFs over e.g. hidden Markov models (HMMs) is their conditional nature, resulting in the relaxation of the independence assumptions that is required by HMMs in order to remain computationally feasible.

We used the CRF++ package with default settings[2].

MBT is a memory-based tagger-generator that generates a sequence tagger on the basis of a training set of labelled sequences, and consecutively can tag new sequences (Daelemans et al., 2003). It has been used to generate POS taggers and various chunkers. MBT can make use of full algorithmic parameters of TiMBL 5.2, a memory-based software package[3].

In our setup, a learner classifies a token from a dialogue (the token under consideration, which we call the *focus token*) in its context of other tokens (the *context tokens*). It depends on internal design how much of a context a sequence learner will consider during classification, we worked with a default token context of 1. For all classifiers we mostly used default settings. It is possible to provide the learners additional information, by means of a vector of features. We discuss our selection of features below.

3.2 Features

Our method for both corpora was to merge all tokens into one single sequence up to a transcribed speaker change. In this way, we preserved a minimum boundary information uniformly for both corpora. In the Monroe dataset a sequence-to-be-chunked on average contains 1.5 DA boundaries, and consists of rather long utterances (e.g., $S4$ and $S5$ in Table 1 would constitute a sequence). In the MRDA dataset, the last two DA units on channel 0 would be merged, as they are transcribed consecutively, but the unit transcribed between $c1$ and $c3$ is regarded as a single-token sequence. By merging the 'utterances' into longer segments of 'turns', we created about 5% less segment boundaries in the MRDA data than in the transcriptions. On average there are 1.8 DA type boundaries in the segments.

The features that we use are straightforward and automatically extractable from the dialogue transcriptions. The majority of these would be internally available from a linearised token stream in a dialogue application as well. Some attributes were derived using some knowledge of transcribed boundaries; this has to do with the fact that although sequence learners can handle a sequence of hundreds of tokens, it is not feasible to feed them an entire dialogue.

Tokens All words were tokenised, dealing with capitalisation, separating and expanding clitics, etc., and subsequently stemmed with a Porter stemmer (Porter, 1980). Apart from taking the word token as a focus feature, we also use the token's part-of-speech tag, automatically obtained by using MBT trained on the Wall Street Journal treebank. We included in the feature vector a context window of 12 left context and six right context elements, both tokens and POS tags. The size of the left context is taken to be the average turn length in tokens, which is estimated 12 for both the Monroe and the MRDA corpus. The context window does not include information contained across the above-explained speaker boundary.

Bag-of-words It has been shown in previous work that redundant encoding of dialogue context may improve the automatic detection of DAs (Lendvai and van den Bosch, 2005). We thus additionally represent lexical context as a bag-of-words (BOW): BOW$left$ contains the lastly uttered 12 words of the current speaker, BOW$leftOth$ contains the most recently uttered 12 words of the speaker that spoke immediately before the focus speaker, and BOW$right$ covers six tokens of right-context for the current speaker only, since it would be incorrect to assume the current speaker to have certainty about what the next speaker will contribute. A threshold on the lexicon size of the BOW has been set to only consider the 200 most frequent word tokens. Note that the BOWs exclude information contained across their own boundaries, and that speaker identity is not encoded.

Silence and overlapping speech For the Monroe data the [SIL] and + markings in the transcriptions were used to derive features. These indicate whether or not an utterance starts or stops with a silence. For the MRDA data, we represented the time elapsed between the previously uttered token in the interaction and the focus token.

3.3 Experimental setup

Our task is to identify in one process for each token in a sequence its DA label, and whether it is a label boundary or not. We represent the DA labels by so-called IOB tags (Tjong Kim Sang and Veenstra,

[2] CRF++ is publicly available at `http://crfpp.sourceforge.net/`

[3] MBT and TiMBL are publicly available at `http://ilk.uvt.nl/`

1999), which is one of the many encoding possibilities. For each DA label a prefix marks whether a token is starting a new DA chunk (`B_<DAtype>`), is inside a DA chunk (`I_<DAtype>`), or outside (`O_<DAtype>`), cf. Table 3. This extended DA label is the class to be guessed by the learners.

	token	Q	S	…	comb.
U	can	I	O	…	I-q
	you	I	O	…	I-q
	see	I	O	…	I-q
	the	I	O	…	I-q
	map	I	O	…	I-q
	have	B	O	…	B-q
	you	I	O	…	I-q
	found	I	O	…	I-q
	it	I	O	…	I-q
S	i	O	I	…	I-s
	can	O	I	…	I-s
	not	O	I	…	I-s
	see	O	I	…	I-s
	it	O	I	…	I-s

Table 3: IOB encoding for questions (Q) and statements (S) in binary classifi cation on the Monroe data.

3.4 Evaluation aspects and metrics

In many previous work on segmentation and classification of dialogue acts, accuracy-based measures such as segmentation and dialogue act error rates have been proposed to asses segmentation and classification performance. Even though these metrics give reasonable insight about performance on the task, higher accuracy or lower error rates do not necessarily imply better performance on DA chunking. Hence we will pay most attention to the traditional measure of information retrieval and chunking: F_1 *score*, a harmonic mean of precision and recall. For comparison with similar work, we additionally report on dialogue act error rate (DER), as described in (Zimmermann et al., 2005): the percentage of misrecognised DAs (i.e., the lower the DER is, the better), where a DA is successfully recognised if both the predicted DA type is correct and the chunk boundaries are successfully predicted. Note that in terms of information retrieval, the DER is none other than the *inverted* DA chunk recall (recall is the proportion of correctly found chunks over the gold-standard amount of chunks). On the token level, we report on the accuracy of predicting the correct IOB tag.

All experiments are carried out separately on the Monroe and on the MRDA datasets. The MRDA dataset allows for multi-class learning, but the Monroe corpus is not annotated with mutually exclusive DAs, yielding over 200, often low-frequent multi-dimensional tags, whose boundaries do not always overlap. Multi-class DA chunking on these data is not straightforward, thus we trained a separate binary classifier for each of the nine occurring DA classes. If we average the results over the classes, we calculate macro averages (in the case of F_1 scores denoted by $F_{1,ma}$). These are in general significantly lower than F micro averages that are calculated proportionally to the amount of each class. We report on both F measures.

4 Experiments and results

On each dataset we run both sequence learners twice: first they have access to the token sequence only, and in a different experiment they can draw on the full feature vector. Additionally, to put the results of CRF and MBL into perspective, we test a baseline method on the DA chunking task, as well as two local classification methods: a Naive Bayes and a k-nearest neighbour approach. The results for Monroe are presented in Table 4 and those for MRDA in Table 5.

4.1 Baselines

A simple majority class baseline is to always guess the majority chunk, which is in both datasets `statement`. This approach labels the beginning of each sequence as `B_statement`, and the rest of the sequence as `I_statement`. We get markedly different scores on the two corpora, since in MRDA the majority of turns include a number of chunks (recall that this material is segmented according to minimal units), whereas in Monroe the segments are typically larger (because the DAMSL annotation scheme allows for assigning multi-level tags to one and the same unit).

When we look to Table 5, we see that for the MRDA dataset this baseline (denoted with *MajChu*) is already rather accurate, (81%), but recalls only a small fraction of the chunks correctly, yielding the relatively low F_1 score (27 points). On the Monroe dataset with separate binary classifiers this labeling clearly is a very bad strategy (8% accuracy, see Table 4), since only one out of the nine binary classifiers has a chance to score at all.

Next, we test powerful local classifiers on the DA chunking task. The naive Bayes classifier is probabilistic and assumes feature independence. It often requires only a small amount of training data to be rather effective. We indeed see that on the Monroe dataset, which contains longer and in a sense more complete utterances, this baseline acquires high accuracy (89%), from only knowing the focus token. When it is provided a relatively large and unorganised additional feature set (recall that the feature vector encodes among others three times 200 bits of contextual bags-of-words), its performance is however dramatically undermined. The same trend can be observed on the MRDA set for the naive Bayes classifier.

Our third baseline is computed by running the IB1 algorithm implemented in the TiMBL package. IB1 is a memory-based learning technique, a direct descendant of the classical *k*-nearest neighbour approach to classification. The number of nearest neighbours used in the experiments was set to nine, and the modified value difference metric was employed in the internal weighting of features. The *k*-nearest neighbours voted on the class using the inverse distance weighting parameter. Note that our sequence learner MBT is also set to employ the IB1 algorithm and the above parameters, thus the differences between a local and a sequential application of the same algorithm are directly comparable. Contrary to the performance of the naive Bayes classifier, IB1's F_1 score improves (on MRDA) or at least remains constant (on Monroe) when it can draw on additional features.

4.2 Sequence learners

A direct comparison between the scores from the two datasets in Tables 4 and 5 may not be informative, due to the differences between these, as explained in Section 2. Nonetheless, we can observe trends within each dataset. The F_1 scores of both sequence learners improve largely over all baselines, indicating that sequential approaches are superior to local classification in the DA chunking task.

CRF's performance is affected in the allFeatures setup to its disadvantage on the Monroe corpus (30 vs 25 F_1,ma), whereas on this material MBT scores identically regardless of the features involved (22 and 23 F_1, ma). The best score is 38 points of micro F score, obtained by the CRF algorithm.

On the MRDA data we see a slight improvement over the token-only experiment for CRF (44 vs 41 F_1, mi). In contrast, MBT's scores seem to weaken on the large feature vector (45 vs 47 F_1, mi).

The two sequence learners work in a rather different way inherently, which explains this divergence. On the smaller dataset (Monroe) CRF performs somewhat better than MBT, especially in the TokenOnly experiment (38 vs 35 F_1, mi), but it is not the case on the large dataset (MRDA), where MBT outperforms CRF in both experimental series.

In general, we see that the magnitude of performance is in the same range for both datasets, despite that it may be more difficult to find a large number of boundaries of short chunks than to identify longer spans of fewer DA type spans. Note that we have much more data from the MRDA corpus, that probably allows some classifiers to be better trained.

Arguably, we set a rather hard task for the learners by limiting the token sequence to material from one speaker only, regardless of own and others' previously uttered tokens, and thereby missing all context that an utterance can have. We deliberately formulated this task, and conjecture that the scores we obtained are in fact out-of-context baseline scores to turn-internal DA chunking, and as such are rather high already. Comparison of our results with previous work cannot be straightforwardly done, due to the differences in creating the sequences that need to be chunked. The obtained DER scores verify the general trend of the sequence learners improving over local classification methods.

We have additionally run experiments to give an impression of the effect of adding more context to the focus token, in the form of the BOW from the immediately previous other speaker (BOW*leftOth*). When splitting down the scores according to DA types, the results indicate that on some DA types there is indeed an improvement over the AllFeatures approach (although not over the TokenOnly experiment), from this additional information. The figures for the two datasets are reported in Table 6 and Table 7.

	tokenOnly				allFeatures			
	Acc	$F_{1,ma}$	$F_{1,mi}$	*DER*	Acc	$F_{1,ma}$	$F_{1,mi}$	*DER*
MajChu	8	3	9	90	8	3	9	90
NBay	89	18	27	75	77	6	6	69
IB1	87	13	22	78	85	13	21	72
CRF	88	30	38	67	84	25	31	70
MBT	86	22	35	68	86	23	33	72

Table 4: Classification performance of nine binary classifiers on the Monroe corpus.

	tokenOnly				allFeatures			
	Acc	$F_{1,ma}$	$F_{1,mi}$	*DER*	Acc	$F_{1,ma}$	$F_{1,mi}$	*DER*
MajChu	81	5	27	78	81	5	27	78
NBay	82	15	16	79	8	7	2	98
IB1	79	1	23	80	83	23	37	61
CRF	83	27	41	65	84	27	44	60
MBT	84	30	47	57	82	29	45	58

Table 5: Classification performance on the MRDA corpus, computed in multi-class learning of seven DA types.

5 Conclusions and future work

In this study we aimed to explore if it is feasible to take a boundary-knowledge-lean approach to jointly segment and label dialogue acts in two corpora. Dialogue processing is dependent on transcribed material, but the representation and segmentation of DA units in dialogue transcriptions is not standardised. Supervised learning of DAs is however dependent on labelled material, where variations of encoding the flow of dialogue supposedly bias the mapping of a dialogue unit to a DA type.

We proposed to refrain from encoding knowledge-based unit boundaries as much as possible, and based DA processing on tokens as basic units. Sequence learning procedures were applied to each token uttered by a speaker, including disfluencies, and a token was classified either as chunk-initial or chunk-internal with respect to a limited set of DA types in the DAMSL, respectively the MRDA annotation scheme.

Two sequence learners, a memory-based tagger and conditional random fields, were trained and tested on the task of segmenting tokens into turn-internal DA chunks. They could draw on a set of straightforward features, or on the token sequence only with a context window of 1. We showed that sequence learning methods are suitable for DA chunking, improving over the results of a chunk majority baseline and local classifiers. The best chunk F_1 score we obtained is 47 points on the transcribed tokens of MRDA spoken discussions, using the MBT sequence tagger in multi-class learning of seven DA types. Note that we used a very strict evaluation metric, the F measure on segmenting and labelling an *IOB*-encoded entire DA chunk.

Our sequence learning methods that performed token-based DA chunking were able to produce comparable F_1, ma scores on rather distinctly transcribed dialogue datasets, both on the MRDA meeting transcriptions and the more traditionally transcribed Monroe scenario dialogues that feature longer turns and a giver-follower dialogue style. Comparing the utility of the lexical token only versus a large bag of straightforward contextual features, we conclude that in our setup lexical items carry the best information for assigning chunk-initial and chunk-internal DA types.

We regard our method as a baseline technique to objectively investigate the role of context in DA chunking. Our plans include explorations on how larger context, including prosodic phenomena, affects performance of sequence learning approaches on DA chunking.

		Agr	Und	Answ	Stat	IList	ISpk	IReq	Conv	Oth
CRF	`TokenOnly`	54	60	23	33	26	21	23	6	26
	`AllFeatures`	45	52	16	29	19	15	16	32	3
	`Token` + `BOW`*leftOth*	47	52	17	30	15	12	12	0	2
MBT	`TokenOnly`	53	58	6	39	13	9	18	0	6
	`AllFeatures`	46	51	11	35	18	15	11	0	17
	`Token` + `BOW`*leftOth*	38	43	10	32	11	8	7	16	14

Table 6: F scores per DA type on the Monroe corpus using different feature sets and sequence learners.

		Backch	Disr	Floorgr	Quest	Statem	X	Z
CRF	`TokenOnly`	69	14	40	23	38	0	5
	`AllFeatures`	68	1	38	20	44	0	21
	`Token` + `BOW`*leftOth*	66	1	28	10	38	0	14
MBT	`TokenOnly`	70	16	39	34	46	0	4
	`AllFeatures`	63	19	40	28	46	0	6
	`Token` + `BOW`*leftOth*	64	18	38	31	42	0	5

Table 7: F scores per DA type on the MRDA corpus using different feature sets and sequence learners.

Acknowledgements

The authors thank Mary Swift, Joel Tetreault, and Amanda Stent for providing the Monroe data, and Elizabeth Shriberg for sharing the ICSI-MRDA dataset. We thank Antal van den Bosch, Sander Canisius, and Erik Tjong Kim Sang for insightful discussions and software help.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch, and Ko van der Sloot. 2003. MBT: Memory based tagger, version 2.0, Reference guide. ILK research group technical report series 03-13, Tilburg.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. TiMBL: Tilburg Memory Based Learner, version 5.1, Reference guide. Technical Report 04-02, ILK, Tilburg University, Tilburg, The Netherlands.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, Institute of Cognitive Science, University of Colorado, USA.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Piroska Lendvai, Antal van den Bosch, and Emiel Krahmer. 2003. Memory-based Disfluency Chunking. In *Proceedings of DISS-03, Disfluency in Spontaneous Speech Workshop*, pages 63—66.

Piroska Lendvai and Antal van den Bosch. 2005. Robust ASR lattice representation types in pragma-semantic processing of spoken input. In *Proceedings of the AAAI Spoken Language Understanding Workshop, SLU-2005*, pages 15–22.

Martin F. Porter. 1980. An algorithm for suffix stripping. In *Program*, *3*(14), pages 130–137.

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

Amanda J. Stent. 2000. The Monroe corpus. Technical Report TR728/TN99-2, University of Rochester, Rochester, UK.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics (EACL)*, pages 173–179, Morristown, NJ, USA. Association for Computational Linguistics.

David R. Traum and Peter A. Heeman. 1996. Utterance Units in Spoken Dialogue. In *ECAI Workshop on Dialogue Processing in Spoken Language Systems*, pages 125–140.

Volker Warnke, Ralf Kompe, Heinrich Niemann, and Elmar Nöth. 1997. Integrated dialog act segmentation and classification using prosodic features and language models. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH-97)*, pages 207–210.

Matthias Zimmermann, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2005. Toward joint segmentation and classification of dialog acts in multiparty meetings. In Steve Renals and Samy Bengio, editors, *MLMI*, volume 3869 of *Lecture Notes in Computer Science*, pages 187–193. Springer.