Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian language

Darinka Verdonik, Matej Rojc, Marko Stabej

Abstract

Speech-to-speech translation technology has difficulties processing elements of spontaneity in conversation. We propose a discourse marker attribute in speech corpora to help overcome some of these problems. There have already been some attempts to annotate discourse markers in speech corpora. However, as there is no consistency on what expressions count as discourse markers, we have to reconsider how to set a framework for annotating, and, in order to better understand what we gain by introducing a discourse marker category, we have to analyse their characteristics and functions in discourse. This is especially important for languages such as Slovenian where no or little research on the topic of discourse markers has been carried out. The aims of this paper are to present a scheme for annotating discourse markers based on the analysis of a corpus of telephone conversations in the tourism domain in the Slovenian language, and to give some additional arguments based on the characteristics and functions of discourse markers that confirm their special status in conversation.

Key words: discourse markers, speech corpora, annotating, conversation, discourse analysis, speech-to-speech translation, spontaneous speech, Slovenian language

1 Introduction

This article is stimulated by problems of speech-to-speech translation technologies that arise from conversational speech phenomena when compared to written language. C-STAR (The Consortium for Speech Translation Advanced Research), the aim of which is to facilitate global cooperation in speech-to-speech translation research, ascertains:

»The fact is humanly spoken sentences are hardly ever well-formed in the sense that they seldom obey rigid syntactic constraints. They contain disfluencies, hesitations (um, hmm, etc.), repetitions (».... so I, I, I guess, what I was saying.«), and false starts (»..how about we meet on Tue.. um.. on Wednesday.....«). Yet put in the context of discussion they are still perfectly understandable for a human listener. A successful speech translation system therefore cannot rely on perfect recognition or perfect syntax. Rather, it must search for a semantically plausible interpretation of the speaker's intent while judiciously ignoring linguistically unimportant words or fragments.« (http://www.c-star.org/main/english/cstar2/; Waibel, 1996)

Many projects developing speech-to-speech translation systems (e.g., Verbmobil – http://verbmobil.dfki.de/, Janus http://www.is.cs.cmu.edu/mie/janus.html, EuTrans – http://www.cordis.lu/esprit/src/30268.htm, Nespole! – http://nespole.itc.it/) had to face the reality of conversational speech. It is commonly noted that conversational speech includes »pauses, hesitations, turn-taking behaviors, etc.« (Kuremtasu et al., 2000), »self-interruptions and self-repairs« (Tillmann, Tischer, 1995), disfluencies such as »a-grammatical phrases (repetitions, corrections, false starts), empty pauses, filled pauses, incomprehensible utterances, technical interruptions, and turn-takes« (Costantini et. al, 2002).

Our intention was not to address all the phenomena of spontaneous speech or to give a surface description of conversational speech phenomena when compared to written text, but to approach the problem of processing conversational speech by a systematic analysis of natural conversation. Analyses of conversation are most common in those fields that are usually encompassed by the common term 'discourse analysis' (see for example Coulthard, 1985; Schiffrin, 1994; Eggins, Slade, 1997; Wood, Kroger, 2000), for example in conversation analysis, sociolinguistics, pragmatics, or systemic functional linguistics, etc. Our analysis was done by considering the needs of speech-to-speech translation. There are many approaches to machine translation, but generally they can be classified either as data-driven methods that are corpus-based and trained on collections of data, or linguistic methods that require an in-depth analysis of sentences and are based on (hand-written) rules (Ueffing et al., 2002). It is the data-driven methods that have shown to be more successful and widely accepted by the machine translation community (Lazzari et al., 2004; Ueffing et al., 2002) and are used in some current projects aimed at improving speech-to-speech translation (e.g., TC-STAR – http://www.tc-star.org/). Data-driven methods are based on large corpora, usually annotated with some linguistic attributes: the most common and basic linguistic attributes are parts of speech, but syntactic or semantic annotations can be included, and at the discourse level there are efforts to annotate corpora with rhetorical relations, anaphoric relations, annotation of temporally sensitive expressions, or expressions of opinions and emotions, etc. Our aim was to specify a category of language elements that are used mostly for communicative purposes, and that could easily be tagged in the corpora needed for developing speech-to-speech translation technology.

When we say communicative purposes, we address to the distinction between the propositional content and the pragmatic functions, as it was specified in some fields of discourse analysis and will be more specifically defined in section 2. This perspective was motivated by the fact that in conversation some parts of text are clearly more important for message content, while others must have more pragmatic, communicative functions in conversation, since their contribution to the message content seems minimal or even zero. Further investigation brings us to the concept of discourse markers as expressions which do not contribute much to the message content. As such they correspond to the C-STAR's suggestion to »search for a semantically plausible interpretation of the speaker's intent while judiciously ignoring linguistically unimportant words or fragments« (see above; http://www.c-star.org/main/english/cstar2/; Waibel, 1996).

There have already been some attempts to annotate discourse markers in speech corpora for use in developing speech technologies or natural language processing (e.g., Heeman et al., 1998; Heeman, Allen, 1999; Miltsakaki et al., 2002). However an overview of the literature on discourse markers

(e.g., Levinson, 1983; Schiffrin, 1987; Redeker, 1990; Fraser, 1996; Blakemore, 1992) shows that there is no consistency on which expressions count as discourse markers, therefore we have to reconsider how to set a framework for annotation. This is especially important for languages such as Slovenian, for which there has been no or little research on the topic of discourse markers (see section 2.4). The aims of this paper are to present a scheme for annotating discourse markers based on the analysis of a corpus of telephone conversations in the tourism domain in the Slovenian language, and to give some additional arguments based on the characteristics and functions of discourse markers that confirm their special status in conversation.

2 Discourse markers

Discourse markers is merely the most popular and common term used to refer to the group of expressions we want to discuss here, however, there are a variety of competing terms used with partially overlapping reference, such as discourse particles, discourse operators, discourse connectives, discourse deixies, pragmatic markers, pragmatic operators, pragmatic particles, etc. (see also Fraser, 1999; Schourup, 1999).

Among the first and the most often cited scholars who drew attention to the group of expressions which indicate relationships between units of discourse were van Dijk, Levinson, Schiffrin, Blakemore, Fraser et al. Van Dijk (1979), for instance, anounced: »In this paper, the pragmatic function of connectives is discussed. Whereas semantic connectives express relations between denoted facts, pragmatic connectives express relations between speech acts. This paper takes a closer look at the pragmatic connectives *and, but, or, so*, and *if*.« A few years later Levinson (1983: 87-88) claims: »/T/here are many words and phrases in English, and no doubt most languages, that indicate the relationship between an utterance and the prior discourse. Examples are utterance-initial usages of *but, therefore, in conclusion, to the contrary, still, however, anyway, well, besides, actually, all in all, so, after all,* and so on. It is generally conceded that such words have at least a component of meaning that resists truth-conditional treatment /.../. What they seem to do is indicate, often in a very complex ways, just how the utterance that contains them is a response to, or a continuation of, some portion of the prior discourse.«

In discourse studies, there has been increasing interest in discourse markers over the past decades, not only in English but many languages worldwide, as can be seen from the number of articles (e.g., Redeker, 1990; Fraser, 1996; Swerts, 1998; Kroon, 1998; Fox Tree, Schrock, 1999; Montes, 1999; Andersen et al., 1999; Archakis, 2001; Matsui, 2001; Schourup, 2001; Norrick, 2001; Vlemings, 2003; Fuller, 2003; Fukushima, 2004; de Klerk, 2004; Tagliamonte, 2005; Dedaić, 2005; Tchizmarova, 2005), special issues (e.g., Discourse Processes, 1997 (24/1); Journal of Pragmatics, 1999 (31/10)), workshops (e.g., Workshop on Discourse Markers, Egmond aan Zee, Nederlands, January 1995; COLING-ACL Workshop on Discourse Relations and Discourse Markers, Montreal, Canada, August 1998), and books (e.g., Schiffrin, 1987; Jucker, Ziv, 1998; Blakemore, 2002) on the subject.

Based on the research of many authors, Schourup (1999) tries to summarize the most prominent characteristics of discourse markers. The following three characteristics are frequently taken to be necessary attributes of discourse markers: 1) connectivity – discourse markers are addressed as items that signal relationships between units of talk; 2) optionality – discourse markers are claimed to be optional (but not redundant!) in two ways: syntactically (the removal of a discourse marker does not alter the grammaticality of its host sentence) and semantically (discourse markers do not enlarge the possibilities for semantic relationship between the elements they associate); 3) non-truth conditionality – discourse markers do not affect the truth conditions of the proposition expressed by an utterance. Schourup (1999) finds the rest of the characteristics less consistent: 4) weak clause association – discourse markers occur either outside the syntactic structure or loosely attached to it; 5) initiality – discourse markers prototypically introduce the discourse segments they mark; 6) orality – most forms claimed to be discourse markers occur primarily in speech; 7) multi-categoriality – discourse markers are heterogeneous with respect to morpho-syntactic categorization (they can be adverbs (*now, anyway*), conjunctions (*and, but, because*), interjections (*oh, gosh*), clauses (*y'know, I mean*) ...).

This short overview of characteristics shows that if we are to search for words or fragments of text that are less important for the message content, discourse markers are a promising category. But it does not reveal much about the functions of discourse markers in conversation; not surprisingly, it is

also insufficient to unambiguously determine the class of discourse markers. In order to better understand how discourse markers function, we will give an overview of the results of various approaches to analysing discourse markers, and finally point to the common features. This way we will try to provide a solid basis for developing a scheme for annotating discourse markers in corpora.

2.1 COHERENCE-BASED APPROACH

One of the first detailed and broadly cited studies on discourse markers was carried out by Schiffrin (1987). She uses the term discourse marker for English expressions *oh, well, and, but, or, so, because, now, then, I mean, y'know* and analyses the usage of these expressions in conversation. Her approach is coherence-based. She proposes a model of coherence in talk, distinguishing five planes of talk: exchange structure (turns, adjacency pairs), action structure (speech acts), ideational structure (semantic units: propositions or ideas), participation framework (social relations between speaker and hearer (e.g., teacher – student), also influenced by the relations of speaker/hearer to talk and ideas, presented in talk), information state (cognitive capacities of speaker/hearer – organization and management of knowledge and meta-knowledge). As a result of her analysis, Schiffrin (1987) concludes that discourse markers are used on these different planes of talk. All markers can indicate more than one plane of talk, however she distinguishes primary planes of use from secondary. The primary plane of use for discourse markers *oh* and *y'know* is information state, the primary plane of use for *well* and *I mean* is participation framework, and the primary plane of use for *and, but, or, so, because, now* and *then* is ideational structure (Schiffrin, 1987, p. 316). She suggests that markers select and then display structural relations on different planes of talk, rather than create such relations. Further she concludes that markers with (referential, semantic, linguistic) meaning, such as conjunctions (*and, but, or, so...*) and time deixies (*now, then*), have their primary functions on ideational planes of talk, and those without meaning, such as lexicalized clauses and particles (*well, oh*), show the reverse tendency. This suggests that »as an expression loses its semantic meaning, it is freer to function in non-ideational realms of discourse« (Schiffrin, 1987: 319). We see this conclusion as an indicator that there may be a broader differences between discourse markers functioning primarily on ideational planes, and all the other discourse markers. In conclusion she proposes additional expressions that should be analysed as discourse markers in some of their uses: *see, look, listen, here, there, why, gosh, boy, say, anyway, anyhow, whatever,* meta-talk such as *this is the point, what I mean is...*

2.2 RELEVANCE THEORY

Within the framework of relevance theory (Wilson, Sperber, 1986), discourse markers are most commonly addressed as discourse connectives. One of the leading authors in this area of research is Diane Blakemore (1992; 2002). According to the relevance-based framework, hearers presuppose that an utterance will have adequate contextual effect for the minimum necessary processing (Blakemore, 1992: 36). Relevance theory developed the distinction between conceptual and procedural meaning. There are two distinct cognitive processes involved in utterance interpretation – linguistic decoding processes, which provide an input to inferential processes (inference), which fill the gap between linguistically encoded representations and conceptual representations. According to this there are

»two ways in which linguistic encoding may act as input to pragmatic inferencing and hence two kind of linguistically encoded meaning: on the one hand, a linguistic expression or structure may encode a constituent of the conceptual representations that enter into pragmatic inferences, while on the other, a linguistic expression may encode a constraint on pragmatic inferences« (Blakemore, 2002: 3-4).

This has become known as the distinction between conceptual vs. procedural encoding/meaning. In this distinction, discourse markers primarily encode procedural meaning. While originally the distinction was envisaged as a cognitive version of the distinctions between truth conditional vs. non-truth conditional, originating in speech act theory, in Blakemore (2002) the author claims this cannot be the case, since there are expressions which encode procedures but contribute to truth-conditional content (e.g., pronouns), and there are expressions which encode concepts but do not contribute to truth-conditional content (e.g., some adverbials). She suggests that discourse markers should be

studied not as operators on the level of discourse, but in terms of their input to cognitive processes underlying successful linguistic communication.

2.3 GRAMMATICAL-PRAGMATIC PERSPECTIVE

Fraser (1990; 1996; 1999) approaches the study of discourse markers from what he himself calls grammatical-pragmatic perspective. One of the basic assumptions of his research is that sentence meaning, the information encoded by linguistic expressions, can be divided up into two separate and distinct parts: the proposition (or propositional content), which represents a state of the world which the speaker wishes to bring to the addressee's attention, and everything else (or pragmatic information): mood markers such as the declarative structure of the sentence, and lexical expressions of varying length and complexity. Propositional content is usually defined by truth-conditionality, i.e. a proposition is a representation of the state of affairs that can be judged either true or not true. Fraser focuses on what is not the proposition and tries to analyse it in terms of different types of signals, which he calls pragmatic markers. Pragmatic markers are, according to Fraser (1996), of four main types: 1) basic markers signal the force of the basic message (e.g., ***I regret*** *that he is still ill*.); 2) commentary markers signal a message with comments on the basic message (e.g., ***Stupidly****, Sara didn't fax the correct form on in time*.); 3) parallel markers signal a message in addition to the basic message (e.g., ***In God's name****, what are you doing now?*); 4) discourse markers signal the relationship of the basic message to the foregoing discourse, more precisely (see Fraser, 1999) between the segment they introduce and the prior segment (e.g., *Jacob was very tired.* ***So****, he left early*.).

Fraser (1999) distinguishes two main classes of discourse markers: 1) markers which relate messages – they relate some aspect of the messages (propositional content, epistemic domain, speech acts) conveyed by segments S2 and S1 (for example, expressions *although, but, conversely, despite (doing) this/that, however, in comparison, in spite of, nevertheless, on the other hand, still, though, whereas, yet,* etc.), and 2) markers which relate topic – they signal a quasi-parallel relationship (adding one more thing, similarity, conclusions, etc.) between S2 and S1 (*above all, also, and, besides, equally, in particular, I mean, likewise, on the top of it all, or, otherwise, too, well, what is more,* etc). Interjections, such as *oh, yeah, yes, no, nope, huh,* etc., are, according to Fraser (1990), not discourse markers.

»While the first class of DMs involved the relationship between aspects of the explicit message of the segment S2 and either an explicit or non-explicit message of the S1, the second class of DMs /.../ involves an aspect of discourse management (Schiffrin's Exchange Structure; Redeker's Sequential Level)« (Fraser, 1999: 949).

Fraser's classification provides a convenient basis for assigning expressions to the category of discourse markers or excluding them, however it has not met with universal acceptance. As Schourup (1999) mentions, his definition has been claimed to be too inclusive, and by virtue of its restriction to relations between successive discourse segments is subject to the criticisms.

2.4 STUDIES ON DISCOURSE MARKERS IN THE SLOVENIAN LANGUAGE

Not many studies of discourse markers exist for the Slovenian language.

Gorjanc (1998) presents a morpho-syntactic typology of connectors, i.e. expressions that usually connect textual segments of various length and establish correlations between clauses and sentences or between a section of the text and its expansion. He examines connectors in scientific texts with respect to their role in the surface construction of the text and in the organization of textual meaning. According to his results, most connectors are conjunctions, but the category also encompasses some relative pronouns, adverbs and particles.

Schlamberger Brezar's study (1998) is based on the theory of the Geneve circle. She briefly presents discourse connectives, further classified into semantic discourse connectives, linking propositions or sequences of propositions, and pragmatic discourse connectives, showing relations between speech acts. On the basis of authentic discourse, she defines the markers of conversational structure (expressions *v bistvu, torej, zdaj, ne, ja, hm, mhm, saj, no...*) on the one hand, and interactional connectives on the other hand, further divided into argumentational connectives,

consecutive connectives, contra-argumentational connectives and re-evaluative connectives. According to Schlamberger Brezar (1998), the first class, i.e. markers of conversational structure, typically lose their lexical meaning. As such, they generally overlap with the discourse markers discussed in this article.

Smolej (2004) focuses her research on particles, i.e. a specific part-of-speech category in the Slovenian grammatical tradition. She is interested in the particles, which do not function on the level of meaning transformation or a precise determination of meaning within parts of a text (meaning modification), but function on the level of textual formation or textual correlation. Particles in the role of textual connectors are defined as textual connective devices, which express meaning and logical relations between sentences or parts of a text.

Pisanski Peterlin is the author of publications on text-organizing metatext in research articles (2002; 2005). She uses the distinction between metatext/metadiscourse and proposition, based on the distinction between truth conditionality and non-truth conditionality. The text-organising metatext she analyses in Slovenian research articles is not completely comparable to discourse markers in conversation as discussed in this article; however, discourse markers can be classified as metadiscourse, and the distinction between metatext/metadiscourse and proposition is also interesting for discourse markers.

2.5 THEORETICAL FRAMEWORK FOR ANNOTATING DISCOURSE MARKERS IN SPONTANEOUS SPEECH

When we try to set a framework for annotating discourse markers in spontaneous speech corpora, we soon find that there is no agreement on what counts as a discourse marker; what is more, some authors even express doubt about whether there is a class of phenomena which can be called discourse markers (e.g., Blakemore, 2002). However, what we find common to the approaches presented above is the acknowledgement that there are two basically different kinds of meaning, communicated by utterances: Schiffrin (1987) distinguishes between the ideational plane on the one hand, and the exchange structure, action structure, participation framework and information state on the other hand; Blakemore (2002) distinguishes between the conceptual and the procedural meaning; Fraser (1996) distinguishes the propositional content from the pragmatic information. Even though these distinctions are not completely parallel, they have a lot in common. Taking into account C-STAR's suggestion to »search for a semantically plausible interpretation of the speaker's intent while judiciously ignoring linguistically unimportant words or fragments« (http://www.c-star.org/main/english/cstar2/; Waibel, 1996), we will look at the expressions which are least important for the ideational plane/conceptual meaning/propositional content, but contribute above all to what we will call pragmatic functions, as expressions of special interest for speech-to-speech translation. Therefore expressions which above all have pragmatic functions will be the center of our interest when annotating discourse markers.

Schiffrin's study (1987) is one of the most extensive, detailed and frequently cited studies of discourse markers based on recorded material of natural conversations, therefore we follow some of her findings. We keep the distinction between the ideational structure and all the other planes of talk – as we pointed out in 2.1, some conclusions in Schiffrin (1987) support the idea that there may be a broader difference between discourse markers functioning primarily on the ideational plane, and discourse markers functioning primarily on all the other planes of talk. A similar distinction is observed by Redeker (1990), who distinguishes between markers of ideational structure and markers of pragmatic structure. Since we are interested in the expressions that function primarily pragmatically and contribute least to the ideational/propositional/conceptual domain, the aim will be to annotate discourse markers that function primarily as pragmatic markers. We take this as the basic theoretical framework for annotating.

We chose the corpus approach to further develop a detailed annotation scheme for discourse markers in the Slovenian language: we collected a corpus of spontaneous conversations, transcribed it, manually annotated the discourse markers in the corpus according to the above guidelines, and analysed the annotated expressions. The results of these analyses are guidelines for further broader annotations of discourse markers in corpora: we can prepare automatic annotation, plan manual correction where necessary, and prepare guidelines for adding new elements to the discourse marker category which we can expect when recording new material, especially in unseen domains. Moreover,

the results of the analyses are also a starting point for further discussions about including discourse marker attributes in the speech corpora used for speech-to-speech translation. This analysis also contributes to a better understanding of some expressions in conversation, which have so far received little interest in Slovenian linguistics.

3 Experiment setup

3.1 DATABASE – THE TURDIS-1 CORPUS

The data was limited to the tourism domain, which has been one of the most common domains of interest in the recent speech-to-speech translation projects (e.g., LC-STAR, EuTrans, Verbmobil, Nespole!, Janus...).

Since tourism in general is too broad as a domain of interest for typical speech-to-speech translation applications, it was further restricted to the following sub-domains:

- telephone conversations in a tourist agency
- telephone conversations in a tourist office
- telephone conversations in a hotel reception

We made two steps to obtain conversations as natural as possible, avoiding most of the problems arising from recording imitated conversations in a studio (unnatural environment, hard-to-motivate speakers, lack of background knowledge for imitating a professional tourist agent (see Verdonik, Rojc, 2006, where the recording of the Turdis database is described in details)), and at the same time to assure in advance the permission of speakers for recording: we contacted professional tourist companies for cooperation, and we enabled the speakers to use the TURDIS recording system in their natural environment, the professional tourist agent at his/her working place, and the potential customer at his/her home, office or anywhere else. Technically this was made possible by using the ISDN card. The TURDIS recording system uses both available ISDN channels. One is used for connecting with an agent and the other for connecting with a caller. Callers do not call a tourist agency directly, instead they call the TURDIS system. The system calls an agent in the selected tourist agency immediately after receiving a call. When both connections are established, the system automatically connects both lines and establishes a direct connection between the caller and the agent. At the same time a recording session on both channels starts. Tourist agents were initially asked for a general permission to record their conversations through the TURDIS system. Callers were contacted individually and asked to make a call; they were mostly employees and students of the University of Maribor. We did not impose many limitations on the topic of conversation since it was already restricted enough by the conversational situation: calls could be made only to two hotel receptions, the local tourist office and four different tourist agencies, all of them in Slovenia. We only encouraged callers to ask for information they might really need or be really interested in, and to rely on their previous experience. All conversations were in Slovenian, which is also the mother tongue of all the callers.

We believe most of the conversations recorded are very natural. Most of the callers stated, that they soon forgot that their call is being recorded because they had to concentrate on the conversation. Only a few callers could not relax, being very nervous throughout the conversation and/or did not know what to say next. The agents were mostly not aware which of their conversations are being recorded, not distinguishing the calls through the Turdis system from all the other calls they normally have at work. There were just few examples when an agent obviously recognized a call through the Turdis system – this was in the first conversations, when the memory of signing a permission to record was still very fresh. Such conversations were not included in the Turdis-1 selection. For most of the conversations in the Turdis-1 selection, we believe that if there was the influence of semi-realistic scenarios, it was strongest at the beginning of the conversations, especially in the part where the caller explains his/her reason for the call, and would fade out from this point on.

Recorded material was orthographically transcribed using the Transcriber tool (http://trans.sourceforge.net/en/presentation.php). We considered some of the EAGLES recommendations (http://www.lc.cnr.it/EAGLES96/spokentx/) and the principles of transcribing BNSI Broadcast News database (Žgank et al., 2004) in transcription. Special tags were included in order to retain information about utterance boundaries when overlapping speech occurs. Background signals (while one speaker is talking, the other participant in conversation uses discourse markers to express

his/her attention, agreement, confirmation, understanding, etc. of what the speaker says, but does not take over the turn to express a new proposition and does not show intention to do so) were not considered as overlapping speech, but were tagged as special overlapping events (e.g., ***[lex=overlap_ja],*** where *lex=overlap* is the description of the event and *ja* (Eng. *yes, yeah*) is the word that was pronounced). For further details of segmentation and transcription, see Verdonik, Rojc, (2006).

At the present, the TURDIS database presents a foundation for future work. It consists of approx. 4.6 hours of recordings (80 conversations), transcriptions include 43,000 words. For the study of discourse markers, we selected 30 conversations and named that data TURDIS-1. The most important reason for making a selection was that we tried to obtain data that would have at least some balance in terms of the speakers' gender, number of agents and callers, length and number of calls to three different types of tourist organizations. Some features could not be controlled when recording. For example we did not know whether the agent answering the phone would be male or female – it turned out there were more female tourist agents, we could not avoid this in the selected data. Similarly, we could not control whether the agent answering the phone would be someone already recorded or a new speaker – for example, the most interesting tourist agency for the callers had only four agents answering the phone, and two of them were recorded very often compared to other agents. A similar situation occurred with the tourist office: conversation with tourist agencies turned out to be longer, the topics were more diverse, tourist agencies were also more interesting for callers compared to hotel receptions or the tourist office. We considered this fact when making the selection for Turdis-1; we tried to achieve a 2 : 1 : 1 ratio (2 for tourist agencies, and 1 for hotels and the tourist office). There were also some "failed" conversations that were discussed above (caller being nervous, etc.), and we did not want to include those in the analysis.

Below, further statistics are provided for the Turdis-1 selection. The total length of the recordings in TURDIS-1 is 106 minutes, the average length of a conversation 3.5 minutes, the number of tokens is 15,717, and the number of utterances 2174. Tables 1 and 2 show more details about the number of utterances, length, number of tokens and number of discourse markers for different types of conversations and different groupings of speakers in the TURDIS-1 database.

Table 1. Statistical data for different types of conversations

	No. of conv.	**No. of utterances**	**Average length**		**Total length**		**No. of discourse markers**
			Minutes	Tokens	Minutes	Tokens	
Tourist agency	14	1077	3.81	555	53.33	7763	1050
Tourist office	8	561	3.51	512	28.1	4094	592
Hotel reception	8	536	3.05	483	24.38	3860	516
Total	**30**	**2174**	**3.54**	**524**	**106.2**	**15,717**	**2158**

Table 2. Statistical data for different groups of speakers (tourist agents, callers; male – M, female – F)

	No. of speakers			**No. of utterances**			**No. of tokens**						**No. of discourse markers**		
							Average			Total					
	M	F	**Total**	M	F	**Total**	M	F	**Total**	M	F	**Total**	M	F	**Total**
Tourist agents	3	17	**20**	221	1071	**1292**	513	468	**475**	1538	7957	**9495**	163	912	**1075**
Callers	14	10	**24**	463	419	**882**	208	332	**259**	2905	3317	**6222**	543	540	**1083**
Total	**17**	**27**	**44**	**684**	**1490**	**2174**	**261**	**418**	**357**	**4443**	**11,274**	**15,717**	**706**	**1452**	**2158**

3.2 METHOD

Discourse markers were manually annotated after the corpus was transcribed. According to our framework for annotating discourse markers, we searched for those uses where an expression contributes least to the propositional content of an utterance. Such expressions were: *ja* (Eng. *yes, yeah, yea, well, I see* – please note that the English expressions are only approximate descriptions in order to help the readers who do not speak the Slovenian language; they are based on the authors' knowledge of English, a Slovenian-English dictionary and the British National Corpus (http://www.natcorp.ox.ac.uk/); the usage of discourse markers is culture-specific and we would need a comparative study in order to specify the English equivalents more accurately), *mhm* (Eng. *mhm*), *aha* (Eng. *I see, oh*), *aja* (Eng. *I see, oh*), *ne?/a ne?/ali ne?/jel?* (no close equivalent in English, rather similar to *right?, y'know, isn't it?,* etc.), *no* (Eng. *well*), *eee/mmm/eeem...* (Eng. *um, uh, uhm*), *dobro/v redu/okej/prav* (Eng. *good, alright, right, okay, well, just*), *glejte/poglejte* (Eng. *look*), *veste/a veste* (Eng. *y'know*), *mislim* (Eng. *I mean*), *zdaj* (Eng. *now*), and background signals (many of the above mentioned expressions: *mhm* (Eng. *mhm*), *aha* (Eng. *I see, oh*), *ja* (Eng. *yes, yeah, yea, I see*), *aja* (Eng. *I see, oh*), *dobro* (Eng. *okay, alright, right*), *okej* (Eng. *okay, alright, right*), and three other expressions: *tako* (Eng. *thus*), *tudi* (Eng. *also*), *seveda* (Eng. *of course*)).

We use the term background signals for events where one speaker is talking and the other participant in conversation uses discourse markers to express his/her attention, agreement, confirmation, understanding, etc., of what the speaker is saying, but does not take over the turn to express a new proposition and does not show intention to do so. The overlapping speech, on the contrary, appears on the turn changing points or when struggling to take-over the next turn/to keep the turn. Background signals were quite frequent. We believe that in speech-to-speech translation technology they should be recognized and treated differently from the regular turn-taking, therefore they were annotated and analysed separately.

In order to confirm the selection and obtain more information on the usage of the annotated expressions, a further analysis were carried out for each expression separately, using the combination of a quantitative and a qualitative approach. The analytical procedure was the following:

1. See if the expression analysed is always a discourse marker or can the same expression also be used as an important element of the propositional content.
2. Count the number of times the expression is used as the discourse marker and as part of the propositional content, if such usage exists.
3. See if there are other (perhaps similar) expressions which are used (more or less) in the same way as the analysed discourse marker. If there are, count how many.
4. Use the conversational analysis method (for a description of this method, see Levinson, 1983, p. 286-287) to analyse the pragmatic functions of the analysed discourse marker.
5. Count the number of uses for the analysed discourse marker at the beginning of an utterance, at the beginning of an utterance with other discourse markers but not in the initial position, as the only word of an utterance, at the end of an utterance, and in the middle of an utterance.
6. See if the analysed discourse marker is used along with other analysed discourse markers, and if there is a typical word order.
7. Count the uses of the discourse markers as background signals and analyse them using the conversational analysis method.

Some of the most interesting results of the analysis are described below.

4 Results of the analysis

The results of the analysis are described in four sections. First, we point to some general findings about each expression: whether the expression is always a discourse marker or can it also be used as an important element of the propositional content, possible differences between these two uses, possible variants of discourse markers, or other similarly used expressions, frequency of use, and other more outstanding characteristics. In the second section we give an overview of the typical positions of discourse markers in an utterance, then we point to the most common collocations of discourse

markers and, finally, try to give an overview of the main pragmatic functions of the analysed discourse markers.

4.1 GENERAL FINDINGS

The expression ***ja*** (Eng. *yes, yeah, yea, well, I see*) is one of the most frequent in our corpus: it is used 323 times plus 226 times as a background signal. *Ja* is traditionally seen as a colloquial particle of agreement or assent, but it can also have pragmatic functions. However, the differences between the two are often hard to define. There are some uses where *ja* is clearly an expression of agreement or assent:

*(1) Ako1: samo Egipt vas[+**SOGOVORNIK_ja**] zanima ? / you are interested only in Egypt[+**OVERLAP_yes**]*
*K25: **ja** / **yes***
K25: eee eee no ali pa ... / um um well or ...

and uses where *ja* is clearly a pragmatic element:

(2) K8: eee koliko koliko pa vam pošiljajo to ? / um how often how often do they send you this (material) ?
K8: ker v eni izmed ta velikih dvoran bi moglo biti / because it should happen in one of the bigger halls
*Ama1: **ja** recimo dvorana Tabor nam ne pošilja programa / **well** the Tabor Hall for example does not send us a program*

Yet there are many uses where it is hard to define whether *ja* is more important as an element of the proposition or as a pragmatic element. In example 3 *ja* expresses K12's assent to what the speaker Aso12 announced he was planning to do, but not as an answer to a question, and it is also repeated twice which is usual for other discourse markers – *mhm* (*mhm*), *aha* (*oh, I see*), *no* (*well*) ...:

(3) Aso12: zdaj konkretno recimo Zaton ne? / now for example Zaton
*K12: **ja ja** Zaton me zanima / **yeah yeah** I am interested in Zaton*

The expressions ***mhm*** (Eng. *mhm*)**, *aha*** (Eng. *I see, oh*) and ***aja*** (Eng. *I see, oh*) are traditionally treated as interjections. When giving the English counterparts, we must warn that the English *oh* shows a greater variety of pragmatic functions than can be assigned to the Slovenian *aha* or *aja* – the use of the English *oh* and the Slovenian *aha* and *aja* overlap only partially. As discourse markers, *mhm, aha* and *aja* function similarly, however, there are differences in use, so they cannot be treated as variants of the same discourse marker. They are often used as background signals, especially ***mhm***: it is used 33 times plus 212 times as a background signal:

(4) Aso1 [overlap]: tudi ta je zelo v redu mislim / this one is also very good I think
*K11 [overlap]: **mhm** tega poznam / **mhm** I know this one*

Aha is used 111 times plus 72 times as a background signal:

(5) K7: ne za dve osebi / no for two persons
*Aso7: za dve osebi **aha** / for two persons **I see***

Aja is rarely used, 4 times plus once as a background signal:

(6) Aso1: žal mi je klime tukaj #ni# / I am sorry there is #no# air-conditioning
*K11: **aja** ni je / **oh** there is none*

No (Eng. *well*) is probably the most typical discourse marker, but it turned out to be less common when compared to other discourse markers (used 51 times) in Slovenian discourse than we had expected on the basis of the uses of the English *well*. Unlike the English *well*, the Slovenian *no* is not *l*, it is traditionally treated as an interjection or a particle. So again we may not treat the English *well* as a complete counterpart. A qualitative analysis showed a great variety of uses for the discourse marker *no*. Here we give two examples (7 and 8), in example 7, *no* introduces a turn where the caller K39 is holding back the agent's enthusiasm for sending him a lot of advertising material, and in example 8 *no*

is repeated many times, introducing a turn where the caller K39 expresses amusement over the fact that the hotel's e-mail is as slow as her own:

(7) Ama1: jaz vam lahko eee čim več tega materiala ne? tudi pošljem da ... / I can um send you as much material as possible so ...
*K39: **no** zdaj ni treba pretiravat / **well** you don't have to exaggerate*

(8) Aha1 [overlap]: ja že pri nas bi / yes we would already [1]
Aha1: [2] bil problem pri pošiljanju [LAUGH] / [2] have a problem while sending (an e-mail) [LAUGH]
*K39: **no no no no** v redu potem smo pa na enaki [.] stopnji / **well well well well** okay then we are on the same [.] level*

Eee (Eng. *um, uh, uhm*) and its variants are traditionally viewed as fillers:

(9) *Ama2: **eee** cene pa zdaj nimam **eee** ker [.] lanske cene ne veljajo ne? / **um** I do not have a price **um** because [.] last year's prices are not valid anymore*

There maybe some hesitation as to whether these language elements should be classified as discourse markers or not, although we can find studies where they are classified as such (Swerts, 1998; Andersen et al., 1999; Montes, 1999). Our framework for searching for linguistic elements which contribute least to the propositional content of an utterance certainly brings fillers, such as the Slovenian *eee,* to our attention. Further our decision that fillers can be treated as discourse markers was supported by the results of our qualitative analysis of their usage, where we concluded that *eee* (and its variants) can be an important instrument in the turn-taking system, that it can point to unexpected events in utterances (such as self-repairing), that it can serve as a signal at the beginning of a turn, or a new topic in a conversation, etc. We transcribed each filler with one word, using characters that would most closely describe its pronunciation according to the tradition of the Slovenian orthography. We perceived seven different transcriptions: by far the most common was *eee* (533 times), there were some variants with the nasal sounds *mmm* (14 times) and *nnn* (7 times), and some exceptional variants described as *eeen* (once), *eeennneee* (once), *eeemmmeee* (once). A slightly different communicative role was noticed for variants ending in the fricatives *h*: *eeeh* (used twice), and *f*: *eeef* (once). Altogether the so-called fillers were used 560 times, which included them among the most frequently used words in our corpus (3% of all words).

A ne?, ali ne? and ***jel?*** are variants of discourse marker ***ne?***, but they are really rare in our corpus (used 4 times altogether). In further analysis they are discussed together with the discourse marker *ne?*. The Slovenian expression ***ne*** shows an important distinction between its function as a discourse marker (no close equivalent in English, rather similar to *right?, y'know, isn't it?,* etc.):

(10) *Api3: mhm sva midva dopoldan govorila[+SOGOVORNIK_ja] **ne?** / mhm we spoke this morning [+OVERLAP_yeah] **right?***

or its function as a negative particle (Eng. *not* and *no*):
- phrases with verb: ***ne** vem (do **not** know), **ne** bi (would **not**), **ne** bo (will **not**), **ne** morem (can **not**)*
*- **ne** tisto ni potrebe (**no** there is no need for that); **ne ne** toliko toliko jih pa ne bo (**no no** there will not be so many)*

The discourse marker *ne?* was used 249 times and is always transcribed with a question mark attached (*ne?*) because it is usually pronounced with a rising intonation. As a negative particle *ne* is used 170 times. While it is usually followed by a verb or used at the beginning of an utterance when it functions as a negative particle, its most typical position as a discourse marker is at the end of an utterance (77%), sometimes also in the middle of an utterance (16%), and very rarely at the beginning of an utterance (3%). The English *right?, y'know* or *isn't it?,* etc. are very approximate description of the Slovenian discourse marker *ne?*. It has no real counterpart in English, in many examples it makes no sense to translate the Slovenian discourse marker *ne?* into English:

*(11) K25: dobro gospa najlepša hvala da ste se tako potrudili **ne?** / okay madam thank you so much for your efforts*

Dobro (Eng. *good, alright, right, okay, well*)***, v redu*** (Eng. *good, alright, right, okay, well*)***, okej*** (Eng. *good, alright, right, okay, well*) and ***prav*** (Eng. *good, alright, right, okay, just*) are homonym discourse markers – their pragmatic functions in discourse are very similar. Their most outstanding

communicative function is that they point to a change of topic or to the closing segment of discourse. *Dobro*, *v redu, okej* and *prav* can also be used as an important element of the propositional content, even though in the TURDIS-1 corpus *okej* (which is a modern colloquial expression borrowed from English) is used only as a discourse marker. The distinction between the uses of *dobro/v redu/prav/okej* as elements of the propositional content and their uses as discourse markers is easy to make. Altogether *dobro/v redu/okej/prav* are used 109 times as discourse markers in the Turdis-1:

(12) K25: ***dobro*** *gospa najlepša hvala da ste se tako potrudili ne? /* ***okay*** *madam thank you very much for your effors*

and 21 times as elements of the propositional content:

(13) K39: ker[+SOGOVORNIK_ja] jim nikoli nič ni ***dobro*** *in vedno* etc. */ because[+OVERLAP_yes] nothing is ever* ***good*** *enough for them and they always* etc.

When *dobro, v redu, okej* and *prav* are used as discourse markers, their position in an utterance is typically initial (40%) or isolated (55%).

Glejte/poglejte (Eng. *look*) are plural imperative forms of the verbs *gledati* (Eng. *to look, to see*) and *pogledati* (Eng. *to look*). The difference is that the first verb is imperfect (progressive) and the second is perfect, but in their function as a discourse marker we found no special difference, except that the progressive form *glejte* was used more often– *glejte* was used 20 times, always as a discourse marker:

(14) Api2: ***glejte*** *do štiri ure je polovična cena /* ***look*** *for four hours it is at half price*

Poglejte was used 7 times as a discourse marker:

(15) Ama1: ja ***poglejte*** *vožnja s splavom eee se prične v mesecu maju / yes* ***look*** *raft rides um begin in May*

and twice as an element of the propositional content:

(16) Ama2: tudi imamo ja ***poglejte*** *pod šport in rekreacija / we have that also yes* ***see*** *under sports and recreation*

Veste (Eng. *y'know*) is a plural indicative form of the verb *vedeti* (Eng. *to know, to be aware, to realize*). Its usage is more diverse than the usage of *glejte/poglejte*. Since it is not used often in the TURDIS-1, the conclusions here merely alert us to its pragmatic functions. In the TURDIS-1 *veste* is used as a discourse marker 13 times:

(17) Ako2: grozni eee eee tako rigorozno kot so pa tu pravila ***veste*** *eee eee / horrible um um so rigorous as rules are here* ***y'know*** *um um*

In more than half of the cases, *veste* as a discourse marker forms a phrase with the interrogative pronouns *kaj* (Eng. *what*), *kje* (Eng. *where*), *koliko* (Eng. *how much*), etc., e.g.,:

(18) K19: ***veste kaj*** *jaz bi se pa pozanimal za tale vaš poslovni klub [.] Piramida /* ***y'know what*** *I am interested in your business club [.] Piramida*

Veste can also be used as an important element of the propositional content – 6 times in the TURDIS-1 corpus:

(19) K3: eee rad bi imel eno eee informacijo če morda ***veste*** *kaj o vožnjo s splavom po Mari() [.] po Dravi / um I need some um information do* ***you*** *perhaps* ***know*** *anything about raft rides through Mari() [.] on the Drava river*

Mislim (Eng. *I mean*) is the first person singular present tense form of the verb *misliti* (Eng. *to think, to believe, to mean*). Its usage as a discourse marker is less clear than that of *glejte/poglejte*. We define *mislim* as a discourse marker only when it can be translated as *I mean* – there are 13 such examples in the TURDIS-1 corpus:

(20) *Ama1: eee[+SOGOVORNIK_mhm] tale cesta **mislim** tako dol ne? se bo spustila[+SOGOVORNIK_ja] in to je za Ribičijo ne? / um[+OVERLAP_mhm] this road **I mean** it will go down[+OVERLAP_yes] and that is for the Ribičija right?*

In all the other uses – 17 in the corpus – we do not define *mislim* as a discourse marker, e.g.,:

(21) *K12: **mislim da** še ne bo eee da ne bo prepozno če še kasneje kaj* etc. */ **I think that** it will not be um that it will not be too late if I later* etc.

Mislim as a discourse marker is rather special in the group of discourse markers, in its pragmatic role it most often points to the text production processes, more specifically, it warns the hearer that the speaker will explain something one more time, etc.

As the last in this group of expressions, we identified the adverb ***zdaj*** (Eng. *now*) as a discourse marker. It is used altogether 143 times, but similarly to *ja* (*yes, yeah, yea, well, I see*), it is very difficult to distinguish between the cases where *zdaj* (*now*) should be annotated as a discourse marker, from those where it is a significant element of propositional content. We tagged 119 cases of *zdaj* as discourse markers. The distinctions are not always clear, however we can find clearly propositional usage on the one hand:

(22) *K23: eee pa se da to nekako da je kakšno informacijo **zdaj** zvem ? / um is it possible that I get some information **now** ?*

and clearly pragmatic usage on the other:

*(23) Aso7: eee **zdaj** hotel Neptun imamo tudi v Tučepih / um **now** we also have the Hotel Neptun in Tučepi*

Yet, in many cases it is very hard to decide which role is more important for *zdaj* (*now*), being an element of the propositional content or a pragmatic element:

(24) K39: eeem treh ali pa štirih Nemcev to zaenkrat še ne vem sss se pravi oni[+SOGOVORNIK_mhm] so pač iz Nemčije[+SOGOVORNIK_mhm] / um three or four German people this I do not know exactly sss so they[+OVERLAP_mhm] are from Germany[+OVERLAP_mhm]
K39: #nikoli# še niso bili v Sloveniji / they have #never# been to Slovenia
*K39: in **zdaj** bi jih ze() pač za takšne štiri pet dni počitnic ki jih bojo imeli v Sloveniji bi jim pač seveda* etc. */ and **now** I would f() for some four five days of vacation they will have in Slovenia I would of course* etc.

When annotating we can still try to distinguish between the usage of *zdaj* as a discourse marker from the usage of *zdaj* as an element of propositional content. However our experience shows that even if there is only one person manually annotating discourse markers, he/she would have difficulties in keeping-up consistency. Therefore, we decided for the time being to annotate all examples of *zdaj* as discourse markers.

Finally, we briefly overview the usage of **background signals**, i.e. discourse markers that the hearer pronounces while the speaker talks, in order to confirm that he/she is listening, that he/she understands, that he/she is (still) interested in the speaker's words, but does not begin a change in turn with it and also does not indicate that he/she is ready to change the turn. There were 554 background signals used in the TURDIS-1 corpus altogether. These were: *aha* (Eng. *I see, oh*) 72 times, *aja* (Eng. *I see, oh*) once, *dobro* (Eng. *okay, alright, right*) 8 times, *ja* (Eng. *yes, yeah, yea, I see*) 213 times and *jaja* (tj. repeated *ja*, pronounced very fast and with no audible pause), 16 times, *mhm* (Eng. *mhm*) 209 times and *mhmmhm* – the same as *jaja* – 3 times, *okej* (Eng. *okay, alright, right*) 3 times, *seveda* (Eng. *of course*) once, *tako* (Eng. *thus*) 23 times, *tudi* (Eng. *also*) 5 times.

Finally we should point to some expressions that we did not classify as discourse markers, even though they function quite similarly in some cases. In our database such examples were *ne vem* (Eng. *I don't know*) and some forms of the verbs of saying.

Ne vem (Eng. *I don't know*) was sometimes used in a way in which its pragmatic functions were very strong, mostly expressing the speaker's attitude to the proposition he/she is introducing. The semantic meaning of *ne vem* in such uses was not literally »not knowing«, but it rather served as an indicator that an example will be given, as in example 36:

(25) *K30: namreč en dan popoldan pa potem naslednji dan v bistvu dopoldan pa še potem [.]* ***ne vem****[+SOGOVORNIK_mhm] do treh popoldan / namely one day in the afternoon and then the next day in the morning and then also [.]* ***I don't know****[+OVERLAP_mhm] till three PM*

Some forms of the verbs of saying, for example *da rečem* (Engl. *so to say*), may be similar to some uses of *mislim* (*I mean*), indicating that the speaker has some trouble searching for appropriate expressions, as in example 37:

(26) *K39: ne tisto ni potrebe zdaj samo zbiram[+SOGOVORNIK_mhm] najprej te [.]* ***da rečem*** *okvirne informacije ki jih potem bi lahko posredovala naprej / no there is no need for that now I merely collect[+OVERLAP_mhm] first this [.]* ***so to say*** *basic information that I could forward*

At present we have not included these expressions in the discourse marker category, mostly because we think their connection to the semantic dimension is very strong. Furthermore, their use in our corpus was not frequent enough to allow a precise analysis. However, how to set the borders of the discourse marker category remains a matter of discussion.

Altogether we annotated 31 different expressions as discourse markers. These expressions were used 2,158 times in the TURDIS-1 corpus as discourse markers, which represents approx. 14% of all the words in the corpus. We believe that this is a rather high percentage. We also pointed to some linguistic and cultural differences in the use of these expressions, which we came across when searching for appropriate translations into English, in order to help our readers better understand the issues discussed.

4.2 TYPICAL POSITIONS OF DISCOURSE MARKERS IN AN UTTERANCE

When analysing the positions of discourse markers in utterances, we distinguish four different positions. The first three positions are at the utterance borders: as the only word of an utterance – the speaker made a pause before continuing his/her turn (position 1), as the first word of an utterance or at the beginning of an utterance, but preceded by one or more discourse markers (position 2), as the last word of an utterance (position 3). We count all other positions as medial (position 4). The positions of background signals were analysed separately.

Table 3 gives the results of the most typical positions for each discourse marker. These results are only for those discourse markers which were used more than ten times. As the most typical, we consider the position in which a discourse marker was used in more than 25% of the cases.

Table 3. The most typical positions in an utterance for the analysed discourse markers

	ja	*mhm*	*aha*	*ne?*	*no*	*eee*	*dobro etc.*	*glejte*	*veste*	*mislim*	*zdaj*
Position 1		+					+				
Position 2	+	+	+		+	+	+	+			+
Position 4						+			+	+	+
Position 3				+	+				+		

According to Table 3 there are only three discourse markers – *eee* (*um*), *mislim* (*I mean*), *zdaj* (*now*) – that are not used typically only in a position at the utterance border. Most of the analysed discourse markers are typically used at the beginning of an utterance, *ne?* (*right?, y'know, isn't it,* etc.) and *no* (*well*) are also typically used at the end of an utterance, and *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*), *mhm* (*mhm*) as the only word of an utterance.

The discourse markers that are typically used at the borders between utterances (positions 1, 2 and 3) – *ja* (*yes, yeah, well, I see*), *mhm* (*mhm*), *aha* (*oh, I see*), *ne?* (*right?, y'know, isn't it?,* etc.), *no* (*well*), *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*), *(po)glejte* (*look*) – were used in these positions 802 times (90%), and in the medial position 91 times (10%). Discourse markers which are (also) typically used in the middle of an utterance (position 4) – *eee* (*um*), *mislim* (*I mean*), *zdaj*

(*now*) – were used in this position 314 times (44%), and in the positions at the borders between utterances (positions 1, 2 and 3) 400 times (56%).

An analysis of typical positions was also carried out for background signals. Their positions were compared to the speaker's utterance. A background signal may be positioned in a pause that the speaker makes, so it does not overlap with the speaker's talk. There were 169 or 31% of such uses for background signals in the corpus. Most of these pauses (159) were between utterances. The rest of the background signals in the corpus – 385 or 70% – overlapped with the speaker's talk. When they overlap, we distinguish between the background signals that overlap with the last word in an utterance that the speaker currently produces (the end of an utterance) – 53 or 9%, and those which overlap with the first word in an utterance that the speaker currently produces (the beginning of an utterance) – 91 or 16%. All the other positions of background signals count as medial – 241 or 44%. We can conclude that approximately half of the uses of background signals are at the borders between utterances that the speaker makes. Table 4 shows the results in terms of percentage. PRIMOŽ

Table 4. Position of background signals according to the utterance that the speaker currently produces

	Pause	Beginning of an utterance	Middle of an utterance	End of an utterance
Background signals	31%	16%	44%	9%

4.3 COLLOCATION OF DISCOURSE MARKERS

163 times (approx. 10% of all instances) the analysed discourse markers were used at the beginning of an utterance, but were preceded by one or more discourse markers. Thus, combinations of discourse markers can be used in collocation. The longest string of this type was:

(27) Ama1: ***ja poglejte eee zdaj*** *v zvezi z Mariborom eee v bistvu mi eee organiziramo samo vodenja /* ***yes look um now*** *concerning Maribor um actually we um organize only guided tours*

When such strings of discourse markers are used, the word order of discourse markers is not totally free (considering the fact that Slovenian is a language with very free word order): *ja* (*yes, yeah, well, I see*) always preceded *glejte* (*look*) and *zdaj* (*now*), but either preceded or followed *eee* (*um*). *Aha* (*oh, I see*) always preceded *zdaj* (*now*), *no* (*well*), *dobro/okej* (*right, okay*), but usually followed *ja* (*yes, yeah, well, I see*). *No* (*well*) followed *aha* (*oh, I see)*, but preceded *zdaj* (*now*). We also noticed that the discourse markers *ja* (*yes, yeah, I see, well*), *aha* (*oh, I see*), *mhm* (*mhm*), *no* (*well*), *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*)*; eee* (*um, uh, uhm*) can be repeated twice or more, but *glejte* (*look*) and *zdaj* (*now*) were never repeated. On the basis of these findings, we tried to define the most typical word order for discourse markers at the beginning of an utterance, when more than one discourse marker is used. This is (we use the »#«sign to point to the discourse markers that can be repeated and the »/« sign to delimit discourse markers which can share a position in a string):

aha#/mhm#/ja# no# dobro#/okej#/v redu#/prav# glejte zdaj

The discourse marker *eee* (*um, uh, uhm*) can also be used in a string of several discourse markers (though not very often – 18 uses or 4%), but it is harder to say whether it has a typical position in such strings. It seems that it can be inserted in any position in the initial string of discourse markers: *eee ja glejte* (*um yes look*), *ja poglejte eee zdaj* (*yes look um now*), *poglejte zdaj eee* (*look now um*).

There were also some discourse markers in the TURDIS-1 corpus that were never used along with other analysed discourse markers: these were *veste* (*y'know*) and *mislim* (*I mean*), as well as those discourse markers that were typically used at the end of an utterance: most commonly *ne?/a ne/ali ne?/jel?* (*right?, y'know, isn't it,* etc.), but also the variants of *ja?* (*yes?*), *dobro?/v redu?* (*right?, okay?*) pronounced with a rising intonation. The discourse markers used at the end of an utterance are usually neither repeated nor used together in collocation, but they do stimulate the hearer to use a background signal or to overtake the turn, often starting it with the discourse markers *ja* (*yes, yeah,*

well, I see), *mhm* (*mhm*), *aha* (*oh, I see*), *dobro* (*right*), etc. In example 26, we see how the uses of *ne?* (*right?*) at the end of an utterance that the speaker Api2 produces are followed by the background signal that the hearer produces:

*(28) Api2: eee se pravi za dva dni da bi imeli **ne? / um so you would have it for two days right?***
*Api2: [**SOGOVORNIK_ja**] / [**OVERLAP_yes**]*
*Api2: zdaj [**+SOGOVORNIK_ja**] odvisno koliko bi bilo tudi nočitev **ne? / now[+OVERLAP_yes**] it depends on how many guests there would be **right?***
*K30: eee **ja** v bistvu trideset nočitev bi bilo ne? / um **yes** actually there would be thirty guests*

4.4 PRAGMATIC FUNCTIONS OF DISCOURSE MARKERS

We used the conversational analysis method to analyse the pragmatic functions of discourse markers. We summed the results of our analyses into four main pragmatic functions that the analysed discourse markers can perform: signalling connections to the propositional content, building relationship between the participants in a conversation, expressing the speaker's attitude to the content of the conversation, organizing the course of a conversation. In this section we give some attention to these conclusions and support them by selected examples from the corpus, which show most of the characteristics discussed. However, we must admit that for most cases it is not possible to say that a discourse marker performs only one of these pragmatic functions.

4.4.1 Signalling Connections to the Propositional Content

We distinguish two directions of signalling and building connections to the propositional content of a conversation: backwards (anaphoric) and forwards (cataphoric). Many of the analysed discourse markers signal anaphoric connections to the previous propositional content. Such discourse markers are: *ja* (*yes, yeah, well, I see*), *mhm* (*mhm*), *aha* (*oh, I see*), *aja* (*oh, I see*), *no* (*well*), *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*), *veste* (*y'know*), *mislim* (*I mean*). In example 27, the speaker Aso1 uses *no* (*well*) in his third turn in order to show that he is continuing the content he started in his first turn in the example:

(29) Aso1: potem [+SOGOVORNIK_ja] zdaj tudi [.] no še še mogoče še boljši je v Osminah [.] v Slanem [+SOGOVORNIK_mhm] / then[+OVERLAP_yes] now also [.] wellan even even better one may be in Osmine [.] in Slano[+OVERLAP_mhm]
Aso1 [overlap]: tudi ta je zelo v redu mislim / this one is also very good I think
K11 [overlap]: mhm tega poznam / mhm I know this one
Aso1 [overlap]: poznate ? / you know this one ?
K11 [overlap]: tega poznam ja / I know this one yes
*Aso1 [overlap]: **no / well** [1]*
K11 [overlap]: ja / yes
Aso1: [2] ta je [.] po mojem vseeno na tem področju še eden [.] [SOGOVORNIK_mhm] tako no [.] najboljših[+LAUGH][+SOGOVORNIK_mhm] eee / [2] this one is [.] I think in this area still one of the [.] [OVERLAP_mhm] well y'know [.] the best[+LAUGH]

Discourse markers that signal cataphoric connections to the propositional content that is to follow, are *(po)glejte* (*look*), *veste* (*y'know*), *zdaj* (*now*). In example 28 the speaker K39 is dictating his e-mail address, and uses the discourse marker *zdaj* (*now*) in his third turn in example 28 to point out that what will follow is the next part of his e-mail address:

(30) K39: ja potem pa pošljite tole kar na ~A pika / yes then send this to ~A dot
Aha1: ~A / ~A
Aha1: to se piše ~A pika normalno ? / this is written normally ~A dot ?
K39: kar ~A pa / just ~A and [1]
K39 [overlap]: [2] potem pika ločilo / [2] then dot the punctuation
Aha1 [overlap]: ja pika ja / yes dot yes
*K39: ~A pika **zdaj** pa moj priimek ki je [priimek] [~P ~R ~I ~M ~E ~K] / ~A dot and **now** my surname which is [surname] [~S ~U ~R ~N ~A ~M ~E]*

4.4.2 Building a Relationship Between the Participants in a Conversation

In the conversations analysed, we noticed that the speaker often checks the hearer's presence, interest in the conversation, understanding, etc., and the hearer confirms his/her presence, interest in the conversation, understanding, etc. The speaker uses the discourse markers *ne?* (*right?, y'know, isn't it,* etc.), *dobro?* (*right?*), *ja?* (*yes?*), *v redu?* (*okay?*) to check the hearer's state, and the hearer uses background signals and the discourse markers *ja* (*yes, yeah, well, I see*), *aha* (*oh, I see*), *mhm* (*mhm*), *dobro* (*good, alright, right, okay*), etc. at the beginning of a new turn (when turn-taking has taken place), to confirm or show his/her state. This type of use also help to build a positive, harmonious relationship between the participants in a conversation. In example 29 we see such a fragment of one of the conversations, where the speaker uses *ne?* (*right?*) to address the hearer, and the hearer uses background signals to respond to the speaker's *ne?* (*right?*):

(31) Ama1 [overlap]: čakajte vam takoj [1] / wait I will tell you [1]
Ama1: [2] povedala / [2] right away
*Ama1: glejte lani **ne?** / look last year **right?***
*Ama1: eee[**+SOGOVORNIK_ja**] je bil tale ma() mali splav do trideset oseb **ne?** / um[+**OVERLAP_yes**] was this a sm() small raft for thirty persons **right?***
*Ama1: [**SOGOVORNIK_aha**] / [**OVERLAP_I see**]*
Ama1: eee je bil nekje okrog štiriinosemdeset tisoč tolarjev / um it was approximately eighty four thousand tolars

4.4.3 Expressing the Speaker's Attitude to the Content of the Conversation

For a few discourse markers, for example *aha* (*oh, I see*), *aja* (*oh, I see*), *no* (*well*), we observed that they can be used to express the speaker's attitude to the content of the conversation. Such uses are not frequent and it seems that this function depends of prosody more than other pragmatic functions discussed here. However, we can not overlook such uses. For example *aha* (*oh, I see*) can express surprise or disappointment, etc., as in example 30, where the speaker K8 is negatively surprised or a little disappointed that he did not get the information he was looking for:

(32) Ama1: žal nimam tukaj nič informacij o tem / I'm sorry I don't have any information about that here
*K8: nimate ? [.] **aha** / you don't have ? [.] **oh***

The discourse marker *no* (*well*) can express dissatisfaction, as in example 31, where it introduces the utterance where the speaker K29 is not completely satisfied with the answer Ako1 has given him:

(33) K29: aha zdaj me pa zanima kako je z eee zdaj nnn eee en dan nazaj je bilo za vizo za Ameriko / I see now I am interested in what is um now em um one day ago there was for a visa for America
K29: kako je s tem zdaj ? / what about this now ?
Ako1: nimamo za Ameriko vize / we don't need a visa for America
*K29: **no** ker zdaj je bilo po radiu nekaj da da [1] / **well** because now there was something on the radio that that [1]*
K29 [overlap]: [2] po novem bomo rabili ... / [2] now we will need ...
Ako1 [overlap]: zaenkrat [1] / for now [1]
Ako1: [2] ni nobene informacije posebne da bi kaj bilo kako drugače / [2] there is no special information that anything is different

4.4.4 Organizing the Course of the Conversation

Organizing the course of the conversation is a very important function of discourse markers. We distinguish between three levels when organizing the course of the conversation: turn-taking, topic switching and disturbances in utterance structure.

Turn-taking is a very delicate system, and conversational analysts who pay particular attention to this subject are surprised to notice that

»less (and often considerably less) than 5 per cents of the speech stream is delivered in overlap (two speakers speaking simultaneously), yet gaps between one person speaking and another starting are frequently measurable in just a few micro-second and they average amounts measured in a few tenths of a second« (Levinson, 1983: 296-297).

We believe that discourse markers contribute much to this fact. Discourse markers, which are most frequently used at the end of an utterance (not necessarily a question, but also an affirmative statement) and which are usually pronounced in a rising intonation (*ne?* (*right?, y'know, isn't it,* etc.), *ja?* (*yes?, dobro?* (*right?*), *v redu?* (*okay?*)), point out that this is the place where the hearer can take over the turn, or even point out that the speaker expects the hearer to take over the turn here. In example 32 the speaker Api3 uses *ne?* (*didn't we*) at the end of his first utterance in the example, indicating that he expects the hearer (K19) to take over the turn now and confirm or deny whether Api3 was right:

(34) K19: eee jaz bi pa se pozanimal za tale poslovni klub ko imate [1] / um I am interested in this business club you offer [1]
K19 [overlap]: [2] zdaj na novo / [2] now the new one
Api3 [overlap]: mhm sva midva [1] / mhm we spoke [1]
Api3: [2] dopoldan govorila[+SOGOVORNIK_ja] ***ne? / [2] this morning didn't we?*** *[+OVERLAP_yes]*
K19: ja pol sem pa jaz bil leteč / yes and afterwards it I was on the go

On the other hand *eee* (*um*) is a typical sign that the speaker has not finished his turn, but will/wants to continue, and *eee* (*um*) and *no* (*well*) indicate that the hearer would like to take over the turn. In example 33 the speaker K23 uses *eee* (*um*) to indicate that he would like to say something, however, the speaker Ako1 does not interrupt what he has already started to say, so K23 does not get a chance to take over the turn immediately:

(35) Ako1: jaz vam [1] / I [1]
Ako1 [overlap]: [2] bom vse [1] / [2] will [1]
K23 [overlap]: ***eee*** *... /* ***um*** *...*
Ako1: [2] poslala če bo pa kaj od tega še za vprašanje pa boste poklical ne? / [2] send you everything and if there is a question you will call right?

Another delicate point in a conversation is the starting or closing of a section. It is delicate

»technically, in the sense that they must be so placed that no party is forced to exit while still having compelling things to say, and socially in the sense that both over-hasty and over-slow terminations can carry unwelcome inferences about the social relationships between the participants« (Levinson, 1983: 316).

One element of achieving agreement about the closing of a conversation is the use of the discourse markers *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*). Levinson (1983) calls them pre-closing items, he mentions *okay, all right, so* for English. Approximately half of the discourse markers *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*) in the TURDIS-1 corpus were used in at the beginning of closing sections. In example 34 the speaker K44 uses *dobro* (*okay*) to introduce the closing section, and the speaker Ane2 uses *v redu* (*alright*) immediately afterwards to express agreement about ending the conversation:

(36) Ane2: se pravi najboljše da se oglasite pa bomo skupaj pogledale ne? / so it would be best for you to come around and we will take a look together right?
K44: ***dobro*** *najlepša [1] /* ***okay*** *thank you [1]*
K44 [overlap]: [2] hvala / [2] very much
Ane2 [overlap]: ***v redu / alright***
Ane2: ja / yes
Ane2 [overlap]: na svidenje / goodbye
K44 [overlap]: na svidenje / goodbye

Discourse markers that indicate disturbances (like repairs or other disfluencies or unexpected changes) in utterance structure are most commonly *mislim* (*I mean*) and *eee* (*um*). In example 35 the speaker uses *eee* (*um*) to indicate the place in the utterance where he will start a repair of some previous segment:

(37) Ako1: zdaj edino če hočete kaj pove() ***eee*** *več vedet če slučajno vejo na ministrstvu za zunanje zadeve v Ljubljani[+SOGOVORNIK_aha] / now if you want to te()* ***um*** *to know more maybe if they know at the Ministry of Foreign Affairs in Ljubljana[+OVERLAP_I see]*

4.5 EXEMPLARY ANNOTATION GUIDELINES FOR THE SLOVENIAN LANGUAGE

From the theoretical framework in section 2 and based on the analysis presented in previous sections, we try to summarize the guidelines for annotating discourse markers in conversations in Slovenian. The annotation can be:

- manual first, and an automatic annotation algorithm can be trained on the basis of a manually annotated database
- automatic first, and manually checked and corrected, where needed.

On the basis of a manually annotated or corrected corpus, automatic annotation of discourse markers can be trained. A similar procedure for annotating discourse markers can be used for any language, only the expressions functioning as discourse markers are different. But we should note that for the Slovenian language the list of discourse markers provided here is not complete; and we give some doubts as to whether discourse markers are a closed category that could be completed.

4.5.1 Guidelines for Manual Annotation

Theoretical guidelines: Discourse markers are expressions in conversation that contribute least to the ideational plane/conceptual meaning/propositional content, and mostly have pragmatic functions:

- they help signal connections to the propositional content,
- they help build a relationship between the participants in a conversation,
- they help express the speaker's attitude to the content of the conversation,
- they help organize the course of a conversation.

Practical issues:

- The content of a message is not affected or is insignificantly affected, if we eliminate a discourse marker from a message.
- Discourse markers are most commonly used at the beginning or at the end of an utterance or isolated from the proposition (as background signals for example), often they are grouped.
- There will always be ambiguous examples of usage where it is hard to define whether an expression functions as a discourse marker or not (for example, some uses of *ja*, *zdaj* or *ne vem*). Decision in such examples should be based on the analysis of pragmatic functions of the expression.

4.5.2 Guidelines for Automatic Annotation

If we decide to do automatic annotation first, before manually checking the corpus, the guidelines are:

- If a turn consists only of the expressions *mhm* (Eng. *mhm*), *aha* (Eng. *I see, oh*), *ja* (Eng. *yes, yeah, yea, I see*), *aja* (Eng. *I see, oh*), *dobro* (Eng. *okay, alright, right*), *okej* (Eng. *okay, alright, right*), *tako* (Eng. *thus*), *tudi* (Eng. *also*), *seveda* (Eng. *of course*)... or a repetition of any of these, the turn-change is unimportant for the content of the conversation and the expressions or a repetition of these expressions should be considered as background signals, i.e. special group of discourse markers.
- Some expressions always function as discourse markers and can be automatically annotated as such without further manual checking. Such expressions incude: *mhm* (Eng. *mhm*), *aha* (Eng. *I see, oh*), *aja* (Eng. *I see, oh*), *no* (Eng. *well*), *eee/eeem/eeen/mmm/nnn/eeemmmeee/eeennneee/eeeh/eeef* (Eng. *um, uh, uhm*)...
- Many expressions can function either as discourse markers or as elements of the propositional content and need to be manually checked. Such expressions include: *ja* (Eng. *yes, yeah, yea, well, I see*), *ne?/a ne?/ali ne?/jel?* (no close equivalent in English, rather similar to *right?, y'know, isn't it?,* etc.), *dobro/v redu/okej/prav* (Eng. *good, alright, right, okay, well, just*), *glejte/poglejte* (Eng. *look*), *veste/a veste* (Eng. *y'know*), *mislim* (Eng. *I mean*), *zdaj* (Eng. *now*)...

When the algorithm for automatic annotation is trained on a manually annotated or manually corrected corpus, distinguishing features that can help improve the performance of such an algorithm include:

- discourse markers are usually positioned at the beginning (*ja* (Eng. *yes, yeah, yea, well, I see*), *dobro/v redu/okej/prav* (Eng. *good, alright, right, okay, well, just*), *glejte/poglejte* (Eng. *look*), *zdaj* (Eng. *now*)) or at the end (*ne?/a ne?/ali ne?/jel?* (no close equivalent in English, rather similar to *right?, y'know, isn't it?,* etc.), *veste/a veste* (Eng. *y'know*)) of an utterance or in isolation (e.g., background signals),
- discourse markers are often grouped or repeated, especially at the beginning of utterances or in isolation.

The expressions listed here are not all the discourse markers which exist in the Slovenian language, but the list will grow with new data. An expert has to read through new data in order to detect other discourse markers, following the guidelines for manual annotation.

4.6 OVERVIEW OF THE ANALYSED DISCOURSE MARKERS

In Table 5 we overview the characteristics of the annotated discourse markers. We give an overview of the most significant pragmatic functions and the most common positions in an utterance only for the expressions that are used in the TURDIS-1 database more than 10 times.

Legend:
The most significant pragmatic functions (see also 4.2.4):
function 1 – signalling connections to the propositional content
function 2 – building a relationship between the participants in a conversation
function 3 – expressing the speaker's attitude to the content of a conversation
function 4 – organizing the course of a conversation

The most common positions in an utterance (see also 4.2.2):
position 1 – in isolation, as the only word of an utterance
position 2 – at the beginning of an utterance
position 3 – at the end of an utterance
position 4 – in the middle of an utterance

Table 5. Overview of the analysed discourse markers

Discourse marker	**The most significant pragmatic functions**	**The most common positions in an utterance**	**Number of cases in the TURDIS-1 corpus**
ja	function 1, function 2, function 4	position 2, background signal	323 + 213 as background signals
jaja			16 as background signals
mhm	function 1, function 2, function 4	position 2, position 1, background signal	33 + 209 as background signals
mhmmhm			3 as background signals
aja			4 + 1 as background signals
aha	function 1, function 2, function 3, function 4	position 2, background signal	111 + 72 as background signals
no	function 1, function 3, function 4	position 2, position 3	51
eee	function 4	position 2, position 4	533
mmm	function 4	position 2, position 4	14
nnn			7
eeen			1
eeennneee			1
eeemmmeee			1
eeeh			2

eeef			1
ne?	function 2, function 4	position 3	249
a ne?			2
ali ne?			1
jel?			1
dobro	function 1, function 2, function 4	position 2, position 1, background signal	46 + 8 as background signals
v redu	function 1, function 2, function 4	position 2, position 1	36
okej	function 1, function 2, function 4	position 2, position 1, background signal	12 + 3 as background signals
prav			4
glejte	function 1, function 2	position 2	20
poglejte			9
veste	function 1, function 2	position 2, position 3	13
mislim	function 1, function 4	position 4	13
zdaj	function 1	position 2, position 4	119
seveda			1 as a background signal
tako			23 as background signals
tudi			5 as background signals
			Total: 2158

5 Discussion

In this paper we tried to provide guidelines for annotating discourse markers, on the basis of an analysis of a corpus of telephone conversations in Slovenian in the tourism domain, and to give some additional arguments based on the characteristics and functions of discourse markers that confirm their special status in conversation. We summarized the guidelines in section 4.5, and tried to confirm the special status of discourse markers by summarizing their characteristics in section 4.6. The analysis shows the most significant characteristics to be the following: discourse markers do not contribute to the content of the message, but mostly perform different pragmatic functions (we defined 4 different functions); most of discourse markers are usually placed at the border between utterances. When we try to translate discourse markers, they are usually not paired one-to-one, so we can say that the use of discourse markers is culture- and language-specific. The frequency of discourse markers in conversations (almost 14% of all the words in our data) indicates that these are important elements of natural conversation.

A speech-to-speech translation system works as a mediator in a natural human-to-human conversation. It interferes with a conversation and influences its flow. When developing speech-to-speech translation, we can try to translate only the information that is semantically important, and eliminate most of the pragmatics of conversation. When following this strategy, the discourse marker tag would point to the group of elements that are not (very) important for the content of a message, so we do not lose important information if we do not translate them.

However, we might want to try to preserve (at least some of) the pragmatics of conversation in speech-to-speech translation. Our analysis showed that discourse markers are important pragmatic elements, language- and culture-specific, and quite frequent. Our study provides the basis for annotating discourse markers in speech. A further comparative study of the use of discourse markers in different languages would give many interesting results and observations on this subject, and for the needs of speech-to-speech translation it could provide some sort of translation scheme or translation procedure concerning discourse markers. We believe that preserving discourse markers in a speech centred translation process would result in a more user-friendly technology.

Another interesting potential topic for future work concerns the prosodic aspect of discourse markers. We did not give much attention to it in our work, since we believe it needs special research

and is therefore beyond the scope of this study. However, there were some indications that discourse markers may be often prosodically marked in an utterance. First, in our data they were often separated from the utterance by a pause: e.g., we indicated a special position in an utterance when a discourse marker is used in isolation, as the only word of an utterance (see tables 4 and 5). Second, the discourse marker *ne?* (*right?, y'know, isn't it?,* etc.), as well as some other discourse markers (e.g., *ja?, dobro?*), were often marked with a rising intonation. Therefore a study of not only the two mentioned but of all prosodic features of discourse markers would be an interesting topic for future research, also important for speech-to-speech translation, since technology strives to preserve the prosodic features of the original utterance in its output.

Acknowledgements

We sincerely thank all the tourist companies that helped us record the conversations for the TURDIS corpus: the **Sonček, Kompas, Neckermann Reisen** and **Aritours** tourist agencies, the **Terme Maribor**, especially the **Hotel Piramida** and the **Hotel Habakuk,** and the **Mariborski zavod za turizem** and its tourist office **MATIC**. We also thank all the tourist agents in these companies whose conversations have been recorded, and all the callers who were ready to use the TURDIS system.

Appendix: Transcription of the examples in this paper

The transcription rules for the examples from the TURDIS-1 corpus in this paper are:

- each caller is identified by the letter *K* and an index number, e.g., *K1*
- each tourist agent is identified by the letter *A*, two lower case letters, indicating the tourist company he works for (e.g., *so* for the *Soncek* tourist agency), and an index number (e.g., *Aso1*)
- the speaker's ID occurs at the beginning of each utterance, or when a turn consists of more than one utterance
- the text of conversations follows a colon sign (:); for overlapping speech the sign *[overlap]* is used between the speaker's ID and the colon sign, for example:
  K1 [overlap]: text
  Aso1 [overlap]: text
- the English translation of each utterance follows a slash sign (/)
- other signs occurring in the examples are:

Sign	**Description**
...	Cut-off utterance.
wor()	Cut-off word
?	Rising intonation.
#word#	Emphasized word.
wo**[:]**rd	Previous phoneme is prolonged.
[.]	Short silence.
Text **[1]**	Utterance continues in the first segment that follows, starting with *[2]*.
[2] *text*	Continuation of the last preceding segment, ending in *[1]*.
text **[P]** *text*	Segment includes two utterances, *[P]* signals the border.
~GMX	Abbreviation is spelled out.
@SI	Abbreviation is pronounced as one word.
[+SOGOVORNIK_ja] / [+OVERLAP_yes]	Background signal *ja* (*yes, yeah*) overlaps with the previous word of the speaker's turn.
[SOGOVORNIK_ja] / [OVERLAP_ja]	Background signal *ja* (*yes, yeah*) is pronounced in a pause that the speaker makes in his talk.
[+LAUGH]	The speaker laughing while pronouncing the previous word.
[LAUGH]	The speaker laughing.

References

1. Andersen, E. S., M. Brizuela, B. DuPuy, L. Gonnermas (1999). Cross-linguistic evidence for the early acquisition of discourse markers as register variables. *Journal of Pragmatics,* 31, 1339-1351.
2. Archakis, A. (2001). On discourse markers: Evidence from Modern Greek. *Journal of Pragmatics,* 33, 1235-1261.
3. Blakemore, Diane (1992). *Understanding utterances*. (Oxford, Cambridge: Blackwell Publishers)
4. Blakemore, Diane (2002). *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. (Cambridge: Cambridge University Press)
5. Constantini, E., S. Burger, F. Pianesi (2002). *NESPOLE!'s multilingual and multimodal corpus*. (Paper presented at 3rd International Conference on Language Resources and Evaluation 2002, LREC 2002, Las Palmas, Spain)
6. Coulthard, M. (1985). *An introduction to discourse analysis*. (London: Longman)
7. Dedaić, Mirjana N. (2005). Ironic denial: *tabože* in Croatian political discourse. *Journal of Pragmatics,* 37, 667-683.
8. Eggins, Suzanne, Slade, Diana (1997). *Analysing Casual Conversation*. (London and Washington: Cassell)
9. Fox Tree, Jean E., Josef C. Schrock (1999). Discourse markers in spontaneous speech: Oh what a difference an *oh* makes. *Journal of Memory and Language,* 40/2, 280-295.
10. Fraser, Bruce (1990). An approach to discourse markers. *Journal of Pragmatics,* 14, 383-395.
11. Fraser, Bruce (1996). Pragmatic markers. *Pragmatics,* 6/2, 167-190.
12. Fraser, Bruce (1999). What are discourse markers? *Journal of Pragmatics,* 31, 931-952.
13. Fukushima, Tatsuya (2004). Japanese continuative conjunction *ga* as a semanetic boundary marker. *Journal of Pragmatics,* 25, 81-106.
14. Fuller, Janet M. (2003). The influence of speaker roles on discourse marker use. *Journal of Pragmatics,* 35, 23-45.
15. Gorjanc, V. (1998). Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija,* XLVI/4, 367–388.
16. Heeman, Peter, Donna Byron, James Allen (1998). Identifying Discourse Markers in Spoken Dialogue. (In *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, Stanford, CA)
17. Heeman, Peter, James Allen (1999). Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. *Computational Linguistics*, **25**(4).
18. Jucker, Andreas H., Yael Ziv (Eds.) (1998). *Discourse Markers: Descriptions and Theory.* (Amsterdam: John Benjamins)
19. de Klerk, Vivian (2004). Procedural meanings of *well* in a corpus of Xhosa English. *Journal of Pragmatics,* 37, 1183-1205.
20. Kroon, Caroline (1998). A framework for the description of Latin discourse markers. *Journal of Pragmatics,* 30, 205-223.
21. Kurematsu, A., Akegami, Y., Burger, S., Jekat, S., Lause, B., MacLaren, V., Oppermann, D., Schultz, T. (2000). *Verbmobil Dialogues: Multifaced Analysis*. (Paper presented at the International Conference of Spoken Language Processing)
22. Lazzari, Gianni, Alex Waibel, C. Zong (2004). *Worldwide ongoing activities on multilingual speech to speech translation*. (Paper presented at Interspeech 2004 - ICSLP, International Conference on Spoken Language Processing, Special Session: Multi-lingual speech-to-speech translation, Jeju Island, Korea)
23. Levinson, Stephen (1983). *Pragmatics*. (Cambridge University Press, Cambridge)
24. Matsui, Tomoko (2001). Semantics and pragmatics of a Japanese discourse marker *dakara* (*so/in other words*): a unitary account. *Journal of Pragmatics,* 34, 867-891.
25. Miltsakaki, E., R. Prasad, A. Joshi, B. Webber (2002). *The Penn Discourse Treebank*. (Paper presented at the Language Resources and Evaluation Conference'04, Lisbon, Portugal)

26. Montes, Rosa Graciela (1999). The development of discourse markers in Spanish: Intejections. *Journal of Pragmatics,* 31, 1289-1319.
27. Norrick, Neal R. (2001). Discourse markers in oral narrative. *Journal of Pragmatics,* 33, 849-878.
28. Pisanski, Agnes (2002). Analiza nekaterih metabesedilnih elementov v slovenskih znanstvenih člankih v dveh časovnih obdobjih. *Slavistična revija,* 50/2, 183-197.
29. Pisanski Peterlin, Agnes (2005). Text-organising metatext in research articles: an English-Slovene contrastive analysis. *Engl. specif. purp. (N.Y. N.Y.),* 24/3, 307-319.
30. Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics,* 14, 367-381.
31. Schiffrin, Deborah (1994). *Approaches to Discourse*. (Cambridge: Blackwell Publishers)
32. Schiffrin, Deborah (1987). *Discourse Markers* (Cambridge: Cambridge University Press)
33. Schlamberger Brezar, M. (1998). Vloga povezovalcev v diskurzu. (In *Jezik za danes in jutri* (pp. 194-202). Ljubljana: Društvo za uporabno jezikoslovje Slovenije)
34. Schourup, Lawrence (1999). Discourse markers. *Lingua,* 107, 227-265.
35. Schourup, Lawrence (2001). Rethinking *well*. *Journal of Pragmatics,* 33, 1025-1060.
36. Smolej, Mojca (2004). Členki kot besedilni povezovalci. *Jezik in slovstvo,* 49/5, 45-57.
37. Swerts, Marc (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics,* 30, 485-496.
38. Tagliamonte, Sali (2005). *So* who? *Like* how? *Just* what? Discourse markers in the conversations of Young Canadians. *Journal of Pragmatics,* 37, 1896-1915.
39. Tchizmarova, Ivelina K. (2005). Hedging functions of the Bulgarian discourse marker *xajde*. *Journal of Pragmatics,* 37, 1143-1163.
40. Tillmann, Hans G., Bernd Tischer (1995). *Collection and exploitation of spontaneous speech produced in negotiation dialogues*. (Paper presented at the ESCA Workshop on Spoken Language Systems, 217-220, Vigsø)
41. Ueffing, N., H. Ney, V. Arranz, N. Castell (2002). *Overview of speech centered translation. LC-STAR*, project report D4.1. http://www.lc-star.com/archive.htm.
42. Verdonik, Darinka, Matej Rojc (2006). *Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems*. (Paper presented at the 5th International Conference on Language Resources and Evaluation, Genoa, Italy)
43. Vlemings, Joeri (2003). The discourse use of French *donc* in imperative sentences. *Journal of Pragmatics,* 35, 1095-1112.
44. Waibel, Alex (1996). Interactive translation of conversational speech. *IEEE Computer,* 29/7, 41-48.
45. Weinrich, Harald (1993). *Textgrammatik der deutschen Sprache*. (Manheim, Leipzig, Wien, Zuerich: Dudenverlag)
46. Wilson, Deirdre, in D. Sperber (1986). *Relevance*. (Cambridge: Cambridge University Press)
47. Wood, Linda A., Kroger, Rolf O. (2000). *Doing Discourse Analysis: Methods for studying action in talk and text*. (Sage Publications, Inc.)
48. Žgank, A., T. Rotovnik, M. Sepesy Maučec, D. Verdonik, J. Kitak, D. Vlaj, V. Hozjan, Z. Kačič, B. Horvat (2004). *Acquisition and annotation of Slovenian Broadcast News database*. (Paper presented at the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal)