

Discourse markers in Slovenian and their applicability for developing speech-to-speech translation technologies

DARINKA VERDONIK
University of Maribor

1. INTRODUCTION

This paper is based on research concerned principally with how linguistic knowledge about conversation could help to improve performance of speech-to-speech translation technologies. Conversations in tourism were the selected domain. One of the important aspects of the research was discourse markers. As I will argue in this paper, the concept of discourse markers could be usefully used in development of speech-to-speech translation technologies. Based on the corpus of recorded telephone conversations in the domain of tourism I define as (some of) the Slovenian discourse markers the expressions: *ja* (Eng. yeah), *mhm* (ah), *aha* (oh), *aja* (oh), *ne?/a ne?/ali ne?/jel?* (right?), *no* (well), *eee/mmm/eeem* (um), *dobro/v redu/okej/prav* (okay/all right), *glejte/poglejte* (look), *veste/a veste* (y'know), *mislim* (I mean), *zdaj* (now), and backchannel signals: *ja* (yeah), *mhm* (ah), *aha* (oh), *aja* (oh), *dobro/okej* (okay/all right), *tako* (that's right), *tudi* (that too), *seveda* (of course). Detailed empirical analysis shows that the functions of these expressions are: signalling connections to propositional content, building relationships between participants in conversation, expressing a speaker's attitude to the content, and negotiating the course of conversation.

Discourse markers have been a fruitful area of research in some fields of linguistics in the last two decades. There have been a number of articles (e.g., Redeker, 1990; Fraser, 1996; Fox Tree, Schrock, 1999; Archakis, 2001; Schourup, 2001;

Vlemings, 2003; Tagliamonte, 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, Netherlands, 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, not only in English but in many other languages as well, eg., Latin (Kroon, 1998), French (Vlemings, 2003), Spanish (Montes, 1999), Catalan (Gonzalez, 2005), Greek (Archakis, 2001), Japanese (Fukushima, 2005), Taiwan Mandarin (Wang et al., 2007), Bulgarian (Tchizmarova, 2005), Croatian (Dedaić, 2005). But concerning the Slovenian language, there have been only a few studies (Gorjanc, 1998; Schlamberger Brezar, 1998; Smolej, 2004; Pisanski Peterlin, 2005) of some of the expressions that can be classified as discourse markers. However, none of them specifically addresses discourse markers and all are based on the analysis of written texts, not spoken discourse.

In developing speech-to-speech translation technologies the need for additional knowledge about natural human conversation appeared. 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’ (Constantini et. al, 2002), and that such phenomena cause difficulties in processing conversational speech. But these are mainly surface observations of conversational speech characteristics; in-depth linguistic analysis of conversation is needed to find explanations and systemic descriptions of conversation phenomena.

In this paper I concentrate on a limited group of expressions that are typically and frequently used in conversation but do not contribute much to its propositional content. I use the widely adopted term ‘discourse markers’, to describe these expressions. Although there may not be complete agreement about the definition of the term and its scope¹, most would concur, I think, that the expressions dealt with here qualify as discourse markers. I begin in any case, by giving an overview of various approaches to the definition of discourse markers, and what they have in common, with the aim of obtaining a category that can be used for pragmatic annotation of speech corpora and thus included in the development of speech-to-speech translation technologies. I analyze the usage of discourse markers in Slovenian conversation following the principles of Conversation Analysis (CA), combining qualitative and quantitative methods, with the aim of gaining more knowledge about the mechanisms that regulate their usage.

The rest of the paper is structured as follows: section 2 presents a short overview of the previous research on discourse markers and specifies the theoretical framework for analyzing them; section 3 provides a short introduction to speech-to-speech translation and an overview of previous attempts to use the discourse markers category in language technologies; section 4 sets out the data

for the analysis and method, and section 5 the results. I here specify the expressions that are used as discourse markers before summarising their pragmatic functions and their typical position in an utterance. In the conclusion I discuss some aspects of the discourse marker category and the potential contribution of the discourse marker tag in speech corpora used for developing speech-to-speech translation technologies.

2. THEORETICAL FRAMEWORK FOR DEFINING AND ANALYZING DISCOURSE MARKERS

As mentioned in the introduction, there have been numerous studies of the range of expressions which have been variously included under the umbrella term of discourse markers in recent years. Among them we find different approaches to, and different definitions of the term 'discourse marker'.

Halliday and Hasan (1976) can be considered among the first to have drawn attention to the lexical and grammatical resources which indicate cohesive relationships between clauses / sentences in text and discourse. They see most of the expressions that were later called discourse or pragmatic markers or connectives as conjunctive elements which represent a special type of cohesive relation with respect to the other three types – lexical organisation, ellipsis and reference – they identify. Conjunction for them is 'rather different in nature from the other cohesive relations'. And further: 'It is not simply an anaphoric [or cataphoric] relation. || Conjunctive elements are cohesive not in themselves but indirectly, by virtue of their specific meaning', and 'express certain meanings which presuppose the presence of other components in the discourse' (Halliday & Hasan, 1976: 226). Halliday & Hasan propose a classification of conjunctions into four categories: additive, adversative, causal, and temporal, and take the words *and*, *yet*, *so* and *then* as typifying these four very general conjunctive relations. Many other expressions can however express each of the relations: e.g., additive can be expressed by *and*, *or*, *nor*, *further*, *furthermore*, *again*, *also*, *moreover*, *what is more*, *similarly*, *on the other hand*, *I mean*, *by the way* etc.; adversative can be expressed by *yet*, *but*, *however*, *though*, *in fact*, *rather*, *any way*, *anyhow* etc.; causal can be expressed by *so*, *thus*, *hence*, *therefore*, *consequently*, *accordingly*, *it follows that* etc.; temporal can be expressed by *then*, *next*, *afterwards*, *after that*, *subsequently*, *soon*, *later* etc.

With regard to interpreting the metafunctional status of conjunctive expressions in discourse, Halliday & Hasan (1976: 239) note concerning temporal conjunction that there is an important distinction between conjunction in the example: 'Next he inserted the key into the lock.' and in the example: 'Next, he was incapable of inserting the key into the lock.' In both examples *next* expresses a temporal relationship. However, in the first example *next* expresses a relation between events; first one thing happens, and then another. It is a relation between meanings in the sense of representations of external reality, therefore the first example 'has to be interpreted in terms of the *experiential* function of language'. But in the second example, the time sequence is in the speaker's organization of his discourse, not of external reality: 'it is a relation between meanings in the sense of representations of the speaker's [...] choice of speech role and rhetorical channel, his

attitudes, judgements and the like'. Therefore the second example 'has to be interpreted in terms of the *interpersonal* function of language'. Halliday and Hasan call the first type of conjunction external, and the second type internal (Halliday & Hasan, 1976: 239-40). This distinction between text-internal and text-external temporal conjunction hints at functions other than textual cohesion, which conjunctive elements can fulfil. The type of interpersonal functions which Halliday and Hasan refer to in their discussion – e.g. expressing a speaker's attitude to a proposition, or indicating his or her speech role – considered 'pragmatic' in other frameworks (see this section, below), are discussed with reference to Slovenian discourse markers in section 5.

Finally, Halliday and Hasan also discuss a number of 'individual items which, although they do not express any particular one of the conjunctive relations identified above, are nevertheless used with a cohesive force in the text' (1976: 267). They refer to these items as continuatives and briefly discuss only six such items: *now*, *of course*, *well*, *anyway*, *surely*, *after all*. Unlike the above, these expressions are more characteristic of spoken than written discourse. Halliday and Hasan note that, when used as continuatives these items are phonetically reduced (i.e., unaccented and with reduced vowel values). In a short discussion of these items (1976: 268-271) they describe the most typical usages of these continuatives: continuative *now* means the opening of a new stage in the communication, *of course* means 'I accept the fact', or rhetorically 'you must accept the fact', *well* introduces a response in dialogue, *anyway* means 'to come back to the point', *surely* means 'am I right in my understanding of what's just been said?', and *after all* means 'what I have just said is reasonable, when everything is taken into account'. In Halliday & Matthiessen (2004: 81) continuatives are defined as 'a small set of words which signal a move in the discourse: a response, in dialogue, or a new move to the next point if the same speaker is continuing. The usual continuatives are *yes no well oh now*.' Further, they say 'essentially [continuatives] constitute a setting for the clause' (Halliday & Matthiessen, 2004: 83). They also note the backchannelling function of continuatives (2004: 154): 'Such items [continuatives] can also function on their own in dialogue, indicating that the listener is tracking the current speaker's contribution.' The SFL concept of the continuative, then, which is associated with some interpersonal functions, can be seen to overlap with what is generally considered as a discourse marker outside the SFL framework; most of the items that will be analyzed in this paper (see section 5 below) would be classified as continuatives in SFL.

The study in this paper was conducted on spoken discourse. In SFL, much less work has been done on spoken discourse than written, and few SFL authors (e.g., Eggins & Slade, 1997; Taboada, 2004) deal (among other things) with markers in spoken discourse. Eggins & Slade (1997: 81-84) discuss expressions that can be interpreted as discourse markers in the framework of adjuncts. Adjuncts for them are 'elements which are additional, rather than essential, to the proposition. They function to add extra information about the events expressed in the core of the proposition' (Eggins & Slade, 1997: 81). Adjuncts may be circumstantial (e.g., *at the moment*, *in the second year*), interpersonal (e.g., *maybe*, *I think*, *I guess*) or textual. Textual adjuncts are sub-classified by Eggins and Slade into:

- conjunctive adjuncts (e.g., *then, next, so, I mean, and*), which link a current clause with prior talk by expressing logical relations of time, cause/consequence, condition, addition, contrast, or restatement, etc.,
- continuity adjuncts (e.g., *oh, well*), which signal that a speaker's clause is coherent with prior talk, without specifying a particular logical relation, and
- holding adjuncts (e.g., *um, ah*), which speakers use to retain the floor while organizing their message.

Only expressions defined as textual adjuncts – some of which ‘indicate a speaker's orientation to the *interactive* continuity of their contribution’ (Eggs & Slade, 1997:84) – are generally interpreted in other frameworks as discourse markers.

A lot of work on discourse markers in spoken discourse can be found outside the field of SFL. One of the first detailed and broadly cited studies was carried out by Schiffrin (1987). Similarly to the SFL approach, she discusses discourse markers within the framework of cohesion. She proposes a model of coherence in talk, distinguishing five planes of talk: exchange structure, action structure, ideational structure, participation framework, information state. 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. Further she concludes that markers with (referential, semantic, linguistic) meaning, such as conjunctions (*and, but, or, so...*) and time deictics (*now, then*), have their primary functions on ideational planes of talk, and those without ideational meaning, such as lexicalized clauses and particles (*well, oh*), have their primary functions on the remaining four planes of talk. 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 difference between discourse markers functioning primarily on ideational planes, and all the other discourse markers.

Within the framework of relevance theory (Wilson & Sperber, 1986), discourse markers are most commonly referred to as ‘discourse connectives’. One of the leading authors in this area of research is Diane Blakemore (1992; 2002). Relevance theory developed the distinction between conceptual encoding/meaning (a linguistic expression or structure encodes a constituent of the conceptual representations that enter into pragmatic inferences) and procedural encoding/meaning (a linguistic expression encodes a constraint on pragmatic inferences). In this distinction, discourse markers primarily encode procedural meaning.

Fraser (1990; 1996; 1999) approaches the study of discourse markers from what he himself calls a ‘grammatical-pragmatic’ perspective. One of the basic assumptions of his research is that sentence meaning, i.e. 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: Mood markers such as the declarative structure of the sentence, and lexical expressions of varying length and complexity.’ (Fraser, 1996: 167) He proposes ‘that this non-propositional part of sentence meaning can be analyzed into different types of signals, what I have called Pragmatic Markers’ (1996: 168-9). He classifies messages and their associated pragmatic markers (which according to

Fraser signal the force of the message, therefore his classification corresponds to messages as well as to pragmatic markers) into four types: basic markers, commentary pragmatic markers, parallel markers, and discourse markers (Fraser, 1996).

For Slovenian there are only a few studies of some elements that could be classified as discourse markers. Gorjanc (1998) presents the morpho-syntactic typology of connectives; Schlamberger Brezar (1998) briefly presents discourse connectives, further classified into semantic discourse connectives and pragmatic discourse connectives; Smolej (2004) discusses particles as connective elements in text, and Pisanski (2002; 2005) presents broader research on text-organizing metatext in research articles. The research reported on here contributes to the exploration of this area of Slovenian language use.

What different, non-SFL approaches to the study of discourse markers have in common is the acknowledgement that there are two basically different kinds of meaning communicated by utterances: Schiffrin (1987) distinguishes between the semantic plane of talk (i.e., the ideational plane), and the pragmatic planes of talk, i.e., the exchange structure, action structure, participation framework and, potentially, information state (Schiffrin, 1987: 24-29); Blakemore (2002) distinguishes between conceptual and procedural meaning and Fraser (1996) distinguishes propositional content from pragmatic information. Even though these distinctions are not completely parallel, they have a lot in common. From this perspective, discourse markers are seen as elements that function primarily on the non-propositional or non-ideational level; in other words, there seems to be common agreement that discourse markers are expressions that are of little importance for the ideational or propositional level, fulfilling mostly what we can call pragmatic functions, and what in SFL would be / are seen as realising interpersonal and textual meanings (see above). This acknowledgement represents the theoretical starting point when I seek to specify items that function as discourse markers in my data.

In analyzing the functions of these markers, I follow the principles of conversation analysis, or CA (ten Have, 1990; Levinson, 1983: 286-332), a research tradition that grew out of ethnomethodology, and which studies the social organization of conversation, or talk-in-interaction, by a detailed inspection of tape recordings and transcriptions made from such recordings. CA is a rigorously empirical approach that 'leaves the researcher with ample room to develop his own best fitting heuristic and argumentative procedures' (ten Have, 1990).

3. DISCOURSE MARKERS IN SPEECH-TO-SPEECH TRANSLATION

Speech-to-speech translation systems incorporate three fields of speech technologies:

- first, we have to transform speech to text, so we use automatic speech recognition;
- then, we automatically translate spoken text from source language (SL) to target language (TL) with a speech-centred translation module;
- finally, we automatically synthesise the TL text so the user can hear it.

All three steps have to work in both directions, from SL to TL and back, so we actually need six modules for one pair of languages, as Figure 1 shows:

FIGURE 1: *The structure of a speech-to-speech translation system for one pair of languages.*

speech recognition (L1) → speech centred translation → speech synthesis (L2)

speech synthesis (L1) ← speech centred translation ← speech recognition (L2)

If a speech-to-speech translation system is to be usable in real-life applications, it has to cope with characteristics of conversational speech like any other application working with human conversation (e.g. spoken dialogue).

Heeman et al. (1998, 1999) discuss an attempt to add a discourse marker tag to a part-of-speech (POS) tagged corpus, claiming that '[t]o understand a speaker's turn of a conversation, one needs to segment it into intonational phrases, clean up any speech repairs that might have occurred, and identify discourse markers' (Heeman & Allen, 1999). Further Heeman et al. (1998) claim 'that discourse markers can be used to help the hearer predict the role that the upcoming utterance plays in the dialog. Thus discourse markers should provide valuable evidence for automatic dialog act prediction.' In their experiment they show 'that discourse markers can be identified very reliably in spoken dialog by viewing the identification task as part of the process of part-of-speech tagging' and that '[t]he identification process can be incorporated into speech recognition, and this leads to a small reduction in both the word perplexity and POS tagging error rate' (Heeman et al., 1998).

Another attempt to annotate discourse markers is the Penn Discourse Treebank. 'With the demand for more powerful NLP (natural language processing) applications comes a need for greater richness in annotation,' claim Miltsakaki et al. They describe 'a new discourse-level annotation project – the Penn Discourse Treebank (PDTB) – that aims to produce a large-scale corpus in which discourse connectives are annotated, along with their arguments, thus exposing a clearly defined level of discourse structure' (2004)².

However, discourse markers are still far from being a well tested and accepted level of corpus annotation. One of the reasons for this might be the fact that discourse markers are not a well defined and unambiguous category. The projects discussed above (Heeman & Allen, 1999) annotate the spoken corpus based on Schiffrin's (1987) research and the project participants' own work (Byron & Heeman, 1997). Following these two references, the list of analyzed discourse markers is very limited: in Byron & Heeman (1997) to the English expressions *and, so, well, oh*, and in Schiffrin (1987) to these and also *but, or, because, now, then, I mean, y'know*. However, other researchers in linguistics and pragmatics mention many other expressions functioning as discourse markers, among which eg. *after all, thus, moreover, however* (Fraser, 1996), and *all right, okay, anyway* (Redeker, 1990). Miltsakaki et al. (2004) on the other hand work mostly with written text corpus and their annotation guidelines are quite different from Heeman's and Allen's (1999): Miltsakaki et al. (2004) count as discourse connectives '(1) all subordinat-

ing conjunctions, (2) all coordinating conjunctions, (3) certain adverbials, and (4) implicit connectives between adjacent sentences’.

4. CORPUS COMPILATION AND CORPUS ANALYSIS PROCEDURE

4.1 THE TURDIS-1 CORPUS

It is very important that real data are used for the analysis in order to be able to study the usage of the chosen set of linguistic elements in conversation. At the time of the research there was no corpus of natural conversation in Slovenian, therefore we had to collect and transcribe our own data. Since this is a very time-consuming and costly procedure, the data were limited. Considering the fact that speech-to-speech-translation is usually developed for a very limited conversational situation, it was appropriate to limit the corpus to one domain of natural conversation. Since tourism has been one of the most common domains of interest in recent speech-to-speech translation projects (e.g., LC-STAR (<http://www.lc-star.com/>), EuTrans (<http://cordis.europa.eu/esprit/src/30268.htm>) and Nespole! (<http://nespole.itc.it/>), which were funded by the European Commission; Verbmobil (<http://verbmobil.dfki.de/overview-us.html>), funded by the German Federal Ministry of Education, Science, Research and Technology, and Janus (<http://www.is.cs.cmu.edu/mie/janus.html>), co-ordinated by Carnegie Mellon University), we chose telephone conversations in travel agencies, tourist offices and hotel receptions.

Recording real conversations can be difficult since speakers have to be notified in advance that their conversation will be recorded. If we want to record telephone conversations between a tourist agent and his customer we could intimidate some customers since not all people are prepared to be recorded. We must also consider that naturalness of conversation is usually affected as soon as a speaker knows that she/he is being recorded. In order to make the conversational situation in recording as natural as possible, we contacted professional tourist companies for cooperation, and we enabled the speakers to use the recording system in their natural environment, professional tourist agents at their workplace and callers at home, in the office or anywhere else. Technically this was made possible by using the ISDN card for recording system. Figure 2 shows how the signal goes from caller to tourist agent and back.



Figure 2: Signal direction in the recording system TURDIS.

We obtained a general permission for recording from tourist agents in the local tourist companies so they were mostly unaware which conversation was being recorded. Callers were contacted individually. We did not set limits on conversation topic since it was already quite restricted by the conversational situation: calls could be made only to two hotel receptions, the local tourist office and four different tourist agencies, all in Slovenia. The callers were encouraged to ask for the information they might really be interested in (more information about the recording process can be found in Verdonik & Rojc, 2006).

Recorded material was orthographically transcribed using the Transcriber tool (<http://trans.sourceforge.net/en/presentation.php>). We used some of the EAGLES recommendations (<http://www.lc.cnr.it/EAGLES96/spokentx/>) and the BNSI Broadcast News database transcription principles (Žgank et al., 2004). We selected 30 conversations from all recorded material, achieving a 2 : 1 ratio (two conversations in tourist agencies for every one conversation in hotels and the tourist office) and named this TURDIS-1. The number of speakers was 44: 20 tourist agents and 24 callers, 17 male and 27 female speakers. 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 15,717, and the number of utterances 2174. Table 1 shows more details about the number and length of conversations.

	NO. OF CONV.	TOTAL LENGTH
Tourist agency	14	53.33 min.
Tourist office	8	28.1 min.
Hotel reception	8	24.38 min.
Total	30	106.2 min.

Table 1. Number and total length of conversations in the TURDIS-1 database.

4.2 PROCEDURE FOR CORPUS ANALYSIS

- (1) Identify and tag discourse marker expressions in the data, following the theoretical framework outlined above, in which discourse markers are defined as expressions that are less important on the ideational or propositional level and fulfil mostly what we have called pragmatic functions.
- (2) Check / verify if the expression analyzed is always a discourse marker or can also be used as an important element of the propositional content, and count the occurrences in each case.
- (3) See if there are other (perhaps similar) expressions which are used in (more or less) the same way as the analyzed discourse marker and count the occurrences.
- (4) Analyze the pragmatic functions of the analyzed discourse marker, following the principles of the CA method.
- (5) Count the discourse marker occurrences at the beginning of an utterance, at the beginning of an utterance with other discourse markers but not in initial position, as the only word of an utterance, at the end of an utterance, and in the middle of an utterance.

(6) Check if the analyzed discourse marker is used along with other analyzed discourse markers, and if there is a typical word order.

(7) Count the uses of discourse markers as backchannel (or background) signals and analyze them using the CA method.

The next section summarizes the most interesting and significant results.

5. RESULTS

Table 1 shows the expressions that were potentially used as discourse markers in our data, and indicates how many instances of these expressions were used as discourse marker or as non-discourse marker, in number of occurrences and in % compared to the total frequency of usage in the corpus.

DISCOURSE MARKER	ENGLISH TRANSLATION ³	NO. OF OCC. AS DM / % OF OCC. AS DM	NO. OF OCC. AS NON-DM / % OF OCC. AS NON-DM
ja*	<i>yes, yeah, yea, well, I see</i>	319 + 245* / 99.30	4 / 0.70
mhm*	<i>mhm</i>	33 + 215* / 100.00	0 / 0.00
aha*	<i>I see, oh</i>	111 + 72* / 100.00	0 / 0.00
aja*	<i>I see, oh</i>	4 + 1* / 100.00	0 / 0.00
ne?, a ne?, ali ne?, jel?	no close equivalent in English, rather similar to <i>right?, y'know, isn't it?, etc.</i>	253 / 59.81	170 / 40.19
no	<i>well</i>	51 / 100.00	0 / 0.00
eee, mmm, eeem ...	<i>um, uh, uhm</i>	560 / 100.00	0 / 0.00
dobro*, v redu, okej*, prav	<i>good, alright, right, okay, well, just</i>	98 + 11* / 82.58	23 / 17.42
glejte, poglejte	<i>look</i>	20 / 90.91	2 / 9.09
veste, a veste	<i>Y'know</i>	13 / 68.41	6 / 31.59
mislilim	<i>I mean</i>	13 / 43.33	17 / 56.67
zdaj	<i>now</i>	119 / 74.84	40 / 25.16
tako*	<i>thus</i>	0 + 23* / 15.65	124 / 84.35
tudi*	<i>also</i>	0 + 5* / 2.84	171 / 97.16
seveda*	<i>of course</i>	0 + 1* / 5.88	16 / 94.12
Total:		2171 / 79.12	573 / 20.88

* Used as a backchannel signal.

Table 1. Expressions potentially functioning as discourse markers, the number of their occurrences in the corpus and % of usage as discourse marker (DM) and non-discourse marker (non-DM) function.

As can be seen from Table 1, the function of discourse marker is much more common than non-discourse marker function for most of the analyzed expressions. Some analyzed expressions always function as discourse marker, and some can function either as discourse marker or as part of the propositional content. For the second type, we can usually easily distinguish both usages, like the use of the expression *glejte/poglejte* (look) in the next two examples, in example 1 as propositional content and in example 2 as discourse marker:

EXAMPLE 1:

*tud mamó ja **poglejte** pod šport in rekreacija / we have as well yes take a **look** at sport and recreation*

EXAMPLE 2:

*ja **poglejte** vožnja s splavom eee se prične v mesecu maju / yes **look** the raft-ride um begins in May*

However, for expressions *ja* (yes, yeah, yea, well, I see) and *zdaj* (now) it is sometimes very difficult to decide whether they should be considered as discourse markers or as propositional content. For example *ja* in example 3 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 (e.g., *mhm* (*mhm*), *aha* (*oh, I see*), *no* (*well*) ...):

EXAMPLE 3:

Aso12: zdaj konkretno recimo Zaton ne? / now for example Zaton;

*K12: **ja ja** Zaton me zanima / **yeah yeah** I am interested in Zaton*

Therefore we can conclude that there is not always a clear line between when an expression functions as a pragmatic element and when as propositional content.

The analysis of the pragmatic functions of discourse markers shows that four main functions are performed:

(1) Signalling connections to the propositional content: there are two possible directions – discourse markers can show connections backwards, to the previous content, i.e. they are anaphoric (*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*)); or forward, to the following content, i.e. they are cataphoric (*(po)glejte* (look), *veste* (*y'know*), *zdaj* (*now*)).

(2) Building a relationship between the conversation participants: there are again two directions: the speaker uses discourse markers to check the hearer's presence, interest in the conversation, understanding, etc. (*ne?* (*right?*, *y'know*, *isn't it*, etc.), *dobro?* (*right?*), *ja?* (*yes?*), *v redu?* (*okay?*)); or the hearer uses discourse markers to confirm his/her presence, interest in the conversation, understanding, etc. (backchannel 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).

(3) **Expressing the speaker's attitude to the content of the conversation:** discourse marker *aha* (*oh, I see*) for example can express surprise or disappointment, etc., *aja* (*oh, I see*) can express surprise for example, *no* (*well*) can express dissatisfaction etc.

(4) **Negotiating the course of the conversation:** I distinguish three levels when negotiating the course of the conversation:

- (a) turn-taking: discourse markers *ne?* (*right?, y'know, isn't it, etc.*), *ja?* (*yes?*), *dobro?* (*right?*), *v redu?* (*okay?*) indicate that this is the place where the hearer can take over the turn; *eee* (*um*) is a typical sign that the speaker has not finished his turn, etc.;
- (b) topic switching: discourse markers *dobro/v redu/okej/prav* (*good, alright, right, okay, well, just*) are important elements in achieving agreement about the closing of a conversation or switching the topic, *no* (*well*) is used when starting a new topic, etc.;
- (c) disturbances in utterance structure: for example when repairs or other disfluencies or unexpected changes appear, discourse markers *mislim* (*I mean*), *eee* (*um*) etc. can be used.

The same discourse marker can realise more of these functions simultaneously (e.g., the same instance of *dobro* (*alright*) can be interpreted as a topic switching signal and as confirmation of understanding), or it can realize different functions in different instances (e.g., in one instance *ne?* (*right?*) can signal that the other participant can take the next turn, and in another instance it can check and attract the hearer's attention).

When analyzing the positions of discourse markers in utterances, I 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). I count all other positions as medial (position 4). Table 2 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. I consider the position in which a discourse marker was used in more than 25% of cases as the most typical.

	<i>ja</i>	<i>mhm</i>	<i>aha</i>	<i>ne?</i>	<i>no</i>	<i>eee</i>	<i>dobro/v redu/ okej/prav</i>	<i>glejte</i>	<i>veste</i>	<i>mislim</i>	<i>zdaj</i>
Position 1		+					+				
Position 2	+	+	+		+	+	+	+			+
Position 4						+			+	+	+
Position 3				+	+				+		

Table 2. The most typical positions in an utterance for the analyzed discourse markers.

Discourse markers that were used at the beginning of an utterance, but preceded by one or more other discourse markers, occurred 163 times (approx. 10% of all instances). Thus, combinations of discourse markers can be used in collocation. When such strings of discourse markers are used, their order 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). 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, I 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 (I use the '#' sign to indicate 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

6. CONCLUSIONS

In this paper I have analyzed a group of Slovenian expressions that contribute very little to the propositional content of conversation and named them discourse markers in accordance with previous research on such expressions in English and other languages. The analysis shows that Slovenian discourse markers are quite diverse in their grammatical form (e.g. interjections, adverbs, verbs etc.) and partly also in the functions they perform. Despite their diversity in form and function, however, they display some systematic commonalities: they do not contribute to the content of the conversation, there is a typical position in an utterance where they are used and a typical word order if more discourse markers are used in a cluster. However, the question of to what extent this is a coherent category, remains open.

Nonetheless, I believe speech-to-speech translation technologies could benefit from annotation of discourse markers in speech corpora. There are three reasons in my view why they could be usefully employed for improving the performance of technologies:

(1) Speech recognition technology in speech-to-speech translation systems includes the so-called language model. Its job is to predict which word in a given string of words is most likely to follow the previous word. In language modelling, POS and other morphological information are often included in corpora used for learning the model in order to improve its performance. As we have seen from our analysis, when an expression functions as a discourse marker, it is more likely to occur at the border between utterances, and when it functions as propositional content, it is more likely to occur within an utterance. If we an-

notate discourse markers in a training corpus and use this information in the language model, it can distinguish these two uses of the same expression. Since discourse markers in natural conversation can be quite frequent (approximately 14% of tokens in our corpus), this may improve the performance of the language model.

(2) Discourse markers often have different translations from the same expressions in propositional content function, e.g., the Slovenian expression *ne* is translated as *right?*, *y'know*, *isn't it*, etc. when used as discourse marker, and as *no* or *not* when used as a negative particle. Both usages are very common. When we use machine translation with a statistical approach (which can be considered the state-of-the-art approach), the translation model is trained on a large parallel corpus. Similarly to language modelling, POS tags are often used to improve the performance of translation models. If we include the annotation of discourse markers as well, this may help the translation model to distinguish better when an expression is used as discourse marker and when as propositional content. So discourse marker annotation may lead to more accurate translations of these expressions.

(3) Just as borders between sentences in written text are important information for both the language model and the translation model, borders between utterances in speech are important information when processing speech. It is important for the language model and the translation model to know when a string of words ends and a new string begins. Besides, we have to segment signals into smaller units that can be processed. However, while in written text it is rather easy to detect borders between sentences, it is not a trivial task to automatically define borders between utterances in speech. As we have seen in our analysis, discourse markers are typically positioned at borders between utterances. By annotating them we provide one more item of information which can be useful for the segmentation of signals into smaller units, appropriate for processing.

Of course each of the above speculations can only be proved or disproved by empirical testing, and this is the subject of further research.

Finally, I would like to briefly consider the issue from a broader perspective. Information on the level of phonetics, morphology, syntax and semantics is commonly included with language resources used for developing language technologies. However, as Mitsakaki et al. (2004) say, 'with the demand for more powerful NLP applications comes a need for greater richness in annotation'. This especially holds for processing dialogue and conversational speech in applications such as speech-to-speech translation, dialogue systems etc. Thus, for example, anaphora resolution has become a common topic in the field of language resources, and we can find attempts to annotate pragmatic elements such as rhetorical relations, self-repairs, expressions of opinion, emotions, etc. In my view, it seems a logical step forward that language and particularly speech technologies will use more and more pragmatic information, integrated from language resources. In this view, annotation of discourse markers is an example of pragmatic information that can be integrated to technologies, in the striving to make them more powerful and user-friendly.

ACKNOWLEDGEMENTS

I sincerely thank all the tourist companies that helped us record the conversations for the TURDIS corpus: the **Sonček, Kompas, Neckermann Reisen** and **Ari-tours** tourist agencies, the **Terme Maribor**, especially the **Hotel Piramida** and the **Hotel Habakuk**, and the **Mariborski zavod za turizem** and its tourist office **MAT-IC**. I 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.

NOTES

1 Various other terms exist (eg., discourse particles, discourse operators, discourse connectives, discourse deixies, pragmatic markers, pragmatic operators, pragmatic particles) to cover similar phenomena.

2 The articles cited here are from conference proceedings that exist only in e-version. There is no paging.

3 The English expressions in brackets are only approximate descriptions to help readers who do not speak the Slovenian language. They are based on the author's knowledge of English, a Slovenian-English dictionary and the British National Corpus (<http://www.natcorp.ox.ac.uk/>).

REFERENCES

- ARCHAKIS, A. (2001), "On discourse markers: Evidence from Modern Greek", in *Journal of Pragmatics*, 33, 1235-1261.
- BLAKEMORE, D. (1992), *Understanding Utterances*. Oxford, Cambridge: Blackwell Publishers.
- BLAKEMORE, D. (2002), *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.
- BYRON, D. K., HEEMAN, P. A. (1997), "Discourse marker use in task-oriented spoken dialog". Paper presented at the 5th *European Conference on Speech Communication and Technology (Eurospeech)*, September, Rhodes, Greece.
- CONSTANTINI, E., BURGER, S., PIANESI, F. (2002), "NESPOLE!'s multilingual and multimodal corpus". Paper presented at the 3rd *International Conference on Language Resources and Evaluation 2002 (LREC 2002)*, May, Las Palmas, Spain.
- DEDAIĆ, M.N. (2005), "Ironic denial: *tabože* in Croatian political discourse", in *Journal of Pragmatics*, 37, 667-683.
- EGGINS, S., SLADE, D. (1997), *Analysing Casual Conversation*. London, Washington: Cassell.
- FOX TREE, J.E., SCHROCK, J.C. (1999), "Discourse markers in spontaneous speech: Oh what a difference an oh makes", in *Journal of Memory and Language*, 40/2, 280-295.
- FRASER, B. (1990), "An approach to discourse markers", in *Journal of Pragmatics*, 14, 383-395.
- FRASER, B. (1996), "Pragmatic markers", in *Pragmatics*, 6/2, 167-190.
- FRASER, B. (1999), "What are discourse markers?", in *Journal of Pragmatics*, 31, 931-952.
- FUKUSHIMA, T. (2005), "Japanese continuative conjunction *ga* as a semantic boundary marker", in *Journal of Pragmatics*, 25, 81-106.
- GONZALEZ, M. (2005), "Pragmatic markers and discourse coherence in English and Catalan oral narrative", in *Discourse Studies*, 7/1, 53-86.
- GORJANC, V. (1998), "Konektorji v slovničnem opisu znanstvenega besedila" (Connectors in the grammatical description of a scientific text). *Slavistična revija*, XLVI/4, 367-388.

- HALLIDAY, M.A.K., MATTHIESSEN, C. (2004), *An Introduction to Functional Grammar: third edition*. London: Edward Arnold.
- HALLIDAY, M.A.K., HASAN, R. (1976), *Cohesion in English*. London, New York: Longman.
- TEN HAVE, P. (1990), "Methodological issues in conversation analysis", in *Bulletin de Méthodologie Sociologique*, 27: 23-51. Available on: http://www2.fmg.uva.nl/emca/mica.htm#N_1.
- HEEMAN, P.A., BYRON, D., ALLEN, J.F. (1998), "Identifying Discourse Markers in Spoken Dialogue". Paper presented in the *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998, Stanford. Available on: <http://www.csee.ogi.edu/~heeman/papers/98-aaais.pdf>.
- HEEMAN, P., ALLEN, J. (1999), "Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog", in *Computational Linguistics*, 25(4).
- JUCKER, A.H., ZIV, Y. (eds.), (1998), *Discourse Markers: Descriptions and Theory*. Amsterdam: John Benjamins.
- KROON, C. (1998), "A framework for the description of Latin discourse markers", in *Journal of Pragmatics*, 30, 205-223.
- KUREMATSU, A., AKEGAMI, Y., BURGER, S., JEKAT, S., LAUSE, B., MACLAREN, V. L., OPPERMANN, D., SCHULTZ, T. (2000), "Verbmobil Dialogues: Multifaced Analysis". Paper presented at the *International Conference on Spoken Language Processing (ICSLP 2000)*. October, Beijing, China. Available on: www.phonetik.uni-muenchen.de/forschung/publikationen/ICSLP2000_KURE.pdf.
- LEVINSON, S. (1983), *Pragmatics*. Cambridge: Cambridge University Press.
- MILTSAKAKI, E., PRASAD, R., JOSHI, A., WEBBER, B. (2004), "The Penn Discourse Treebank". Paper presented at the *4th Language Resources and Evaluation Conference'04 (LREC'04)*, May, Lisbon, Portugal.
- MONTES, R.G. (1999), "The development of discourse markers in Spanish: Intejctions", in *Journal of Pragmatics*, 31, 1289-1319.
- PISANSKI, A. (2002), "Analiza nekaterih metabesedilnih elementov v slovenskih znanstvenih člankih v dveh časovnih obdobjih' (An analysis of selected categories of metatext in Slovene research articles in two periods)", in *Slavistična revija*, 50/2, 183-197.
- PISANSKI PETERLIN, A. (2005), "Text-organising metatext in research articles: An English-Slovene contrastive analysis", in *English for Specific Purposes*, 24/3, 307-319.
- REDEKER, G. (1990), "Ideational and pragmatic markers of discourse structure", in *Journal of Pragmatics*, 14, 367-381.
- SCHIFFRIN, D. (1987), *Discourse Markers*. Cambridge: Cambridge University Press.
- SCHLAMBERGER BREZAR, M. (1998), "Vloga povezovalcev v diskurzu" (The role of connectors in spoken discourse). ŠTRUKELJ, I. (ed.), *Jezik za danes in jutri*. Ljubljana: Društvo za uporabno jezikoslovje Slovenije. 194-202.
- SCHOURUP, L. (1999), "Discourse markers", in *Lingua*, 107, 227-265.
- SCHOURUP, L. (2001), "Rethinking well", in *Journal of Pragmatics*, 33, 1025-1060.
- SMOLEJ, M. (2004), "Členki kot besedilni povezovalci" (Particles as textual connectors), in *Jezik in slovnstvo*, 49/5, 45-57.
- TABOADA, M.T. (2004), *Building Coherence and Cohesion: Task-oriented Dialogue in English and Spanish*. Amsterdam, Philadelphia: John Benjamins.
- TAGLIAMONTE, S. (2005), "So who? Like how? Just what? Discourse markers in the conversations of Young Canadians", in *Journal of Pragmatics*, 37, 1896-1915.
- TCHIZMAROVA, I.K. (2005), "Hedging functions of the Bulgarian discourse marker *xajde*", in *Journal of Pragmatics*, 37, 1143-1163.
- TILLMANN, H.G., B. TISCHER (1995), "Collection and Exploitation of Spontaneous Speech Produced in Negotiation Dialogues". Paper presented at the *ESCA Workshop on Spoken Language Systems*, 217-220, Vigso, Denmark. Available on: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.5121>.
- VERDONIK, D., ROJC, M. (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 (LREC'06)*, May, Genoa, Italy.
- VLEMINGS, J. (2003), "The discourse use of French *donc* in imperative sentences", in *Journal of Pragmatics*, 35, 1095-1112.
- WANG, Y.F., TSAI, P.H., LING, M.Y. (2007), "From informational to emotive use: *meiyou* ('no') as a discourse marker in Taiwan Mandarin conversation", in *Discourse Studies*, 9/5, 677-701.
- WILSON, D., D. SPERBER (1986), *Relevance*. Cambridge: Cambridge University Press.
- ŽGANK, A., ROTOVNIK, T., SEPESY MAUČEC, M., VERDONIK, D., KITAK, J. VLAJ, D. HOZJAN, V. KAČIČ, Z. HORVAT, B. (2004), "Acquisition and annotation of Slovenian Broadcast News database", in Paper presented at the *4th International Conference on Language Resources and Evaluation (LREC'04)*, May, Lisbon, Portugal.