Iztok Kosem
Trojina, Institute for Applied Slovene Studies*

UDK 811.163.6'322:808.5:81'373

Darinka Verdonik
University of Maribor**

KEY WORD ANALYSIS OF DISCOURSES IN SLOVENE SPEECH: DIFFERENCES AND SIMILARITIES

1 INTRODUCTION

This study addresses the differences and similarities between different types of speech. There is a great deal of research available that compares linguistic characteristics (lexical and grammatical) of different discourses, registers or genres. The comparisons are more often made at a specific level, i.e. in terms of a particular feature in the discourse, rather than at a general level. One well-known general comparison comprised a corpus-based study of four major registers of English (conversation, fiction, newspapers and academic writing), and was conducted by Biber et al. (1999). Although this comprehensive study considered a spoken register, it described linguistic features of the spoken register in comparison with three written registers; therefore, it provides no information on internal variety of speech.

One of the main obstacles to wider research into internal variety of speech is the lack of large and balanced spoken corpora. Academic language, especially English, seems to be an exception – the availability of spoken corpora such as the British Academic Spoken English (BASE) corpus,[1] the Michigan Corpus of American Spoken English (MICASE),[2] and the spoken part of the corpus used in the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) project has resulted in considerable research into the variety of spoken academic registers (e.g. Nesi 2001; Biber et al. 2002; Biber 2006; Louwerse et al. 2008; Lee 2009). While many of these studies report on internal variety in academic speech in terms of individual linguistic features, one of the important findings of the studies on the T2K-SWAL project is that "all [academic] spoken registers are similar in their typical linguistic characteristics" (Biber 2006:223).

* *Author's address:* Trojina, Zavod za uporabno slovenistiko, Partizanska cesta 5, 4220 Škofja Loka, Slovenia. E-mail: iztok.kosem@trojina.si

** *Author's address:* Fakulteta za elektrotehniko, računalništvo in informatiko, Laboratorij za digitalno procesiranje signalov, Smetanova ulica 17, 2000 Maribor, Slovenia. E-mail: darinka.verdonik@uni-mb.si

[1] http://www2.warwick.ac.uk/fac/soc/celte/research/base

[2] http://quod.lib.umich.edu/m/micase/

For Slovene, one can observe a situation similar to other languages – speech has attracted considerable attention in the past decade; however, few studies have investigated the differences and similarities between different spoken discourses (e.g. Verdonik 2010; Verdonik et al. 2008). Research has mainly focussed on speech in general (e.g. Kranjc 1999; Smolej 2007; Volk 2007), one discourse type (e.g. Krajnc 2005), or a particular linguistic feature (e.g. Verdonik 2007). The recently created first reference spoken corpus of Slovene, called GOS (presented in more detail in the next section), is an important resource that should help researchers fill this gap in research on variety in Slovene speech.

This paper makes the first step towards filling this gap by examining the distinguishing lexical items of the five different discourse types of the GOS corpus, and identifying lexical characteristics shared by different discourses. We employed the key word extraction method to identify statistically significant lexical items. An important finding is that the presence or absence of a particular word class in the key word list can be a good indicator of the type of spoken discourse, or discourses.

2 DATA AND METHOD

The data used for the analysis were from the Slovene reference speech corpus GOS (Verdonik/Zwitter Vitez 2011), consisting of 1,032,775 words or 120 hours of recordings. It contains speech events from five different discourse types with different channels (Table 1). An important characteristic of the corpus, ensuring a good comparability of different discourse types, is that the majority of the recordings consist of spontaneous speech (as opposed to read speech).

Discourse type	Channel	Number of tokens	Totals	Percentage
Classes		162,750	162,750	16%
Media - informative	radio	94,536	196,799	19%
	TV	102,263		
Media - entertainment	radio	123,152	228,765	22%
	TV	105,613		
Official	phone	33,484	153,471	15%
	personal communication	119,987		
Private	phone	68,083	290,990	28%
	personal communication	222,907		

Table 1: Discourse types in the GOS corpus.

We used a well-established method of key word extraction (Scott 1997; 1998; 2000; 2001a; 2001b; 2002) – applied by researchers such as Tribble (2000), Toolan (2004), Johnson and Esslin (2006), and Duguid (2010) – to identify typical lexical items of each discourse, and observe characteristics shared by different discourses. Key words are words that occur unusually frequently in a text or corpus of texts compared to their frequency in a reference corpus (i.e. the difference between their frequency in the study corpus and their frequency in the reference corpus is statistically significant).

A very important aspect of key word analysis is the selection of a reference corpus. A reference corpus is normally much larger than the text or sub-corpus under analysis. In terms of content, Scott and Tribble (2006: 58) suggest that the reference corpus "should be an appropriate sample of the language which the text we are studying (the 'node-text') is written in." The GOS corpus was the only corpus of Slovene speech that met these conditions for our study. It is not common for the reference corpus to be obtained from the corpus under analysis when employing the method of key word extraction. However, we justify the use of this method for this study as other corpora of Slovene speech tend to be smaller or more specialized, for example Broadcast News Speech Database (Žgank et al. 2004; Žibert/Mihelič, 2004) or the Sloparl corpus (Žgank et al. 2006), which contains transcriptions of parliamentary debates, and were thus unsuitable for our purposes. Each discourse sub-corpus under analysis was excluded from the GOS to create its respective reference corpus, to improve key word extraction results (Jeon/Choe 2009) – consequently, five different reference corpora were used, one for each discourse sub-corpus of GOS.

The list of key words for each discourse was extracted using the WordSmith Tools 5.0 (Scott 2008) computer program. In order to be able to conduct this statistical analysis, we first prepared word lists for discourse sub-corpora and for reference corpora. Word lists were prepared for word forms (as opposed to lemmas), as this was expected to give more reliable results; in addition, by focussing on word forms, some indication of the relevance of the word's grammatical characteristics, such as case, gender and number could be deducted directly from the key word list.

The statistical test used for key word extraction was loglikelihood ($p<0.001$). The minimum frequency of a word from a particular discourse to be listed a key word was initially set to 3, in order to exclude hapax legomena and very rare words. The key word lists obtained differed in the number of words they contained: there were 2566 key words for classes discourse, 2992 for media-informative discourse, 2511 for media-entertainment discourse, 1091 for official discourse, and 1550 for private discourse. Because the number of words in each list was very large, we decided to focus on top key words in each list (up to 10 %), but rather than taking the top X % of words on each list, we set a keyness threshold of 50, as this enabled a better comparability of the analysed key lexical items. This resulted in the final lists of items for our analysis: 220 for classes, 213 for media-informative, 133 for media-entertainment, 51 for official, and 131 for private discourse.

The qualitative part of the analysis involved examining concordances of each word on the final lists, eliminating words that featured due to their high frequency of use in just one discourse or by just one speaker, and identifying the topic-bound words. Also, corpus noise items, such as speaker IDs, # symbol, etc., were removed. The remaining words were categorized according to word classes. During manual examination of corpus concordances we noted the most common, outstanding, or non-standard usages (including the most common phrases that contain a key word). We also attempted to establish a connection to the pragmatic aspects of word usage, wherever possible.

3 RESULTS OF THE ANALYSIS

The results are presented as a summary of key words according to standard word classes (nouns, verbs, adjectives etc.). However, during the manual examination of concordances, it became evident that some words could also be grouped on pragmatic aspects (e.g. discourse markers, markers of approximation, fillers, greetings etc.), and that several key words often featured in a phrase; therefore the phrase, rather than just the word, could be assigned to a group.

3.1 Classes discourse

The most common group of key words in classes discourse is verbs. They indicate the common actions that are going on in the class: a teacher explains ((*se) pravi* 'it means', *predstavlja* 'it represents'), introduces new terminology (with verbs such as: *imenujemo* 'we name', *rečemo* 'we say'), categorizes (*spada* 'it belongs'), instructs pupils (*napišite* 'write down', *preberi* 'read'), repeats what has been told, for improved recall (*smo rekli* 'we said') etc. Plural is the most common verb form: in the second person when the teacher refers to pupils, and in the first person when the whole class cooperates in the same action. The verb *prosim* 'please', which is located very high on the key word list, indicates politeness conventions. Direct imperative forms are commonly found (e.g. *preberi* 'read'). Among other typical verb forms are future forms of the auxiliary *biti* 'be'.

Another well-represented group of key words is pronouns, especially those used to ask questions: *kaj* 'what', *katere* 'which', *(iz/od) česa* '(from) what', *kdo* 'who' (referring to pupils: *kdo manjka* 'who's missing'). Demonstrative pronouns, such as *tale* 'this one', are also typical and are used as deictics, to point to the subject of discussion.

Nouns are also often found among key words in classes, but are mostly related to a particular discourse topic (e.g. *povedek* 'predicate', *polmer* 'radius', *verjetnost* 'probability'). There are few typical topic-free nouns, like *del* 'part', *učitelj/-ica* 'the teacher' etc.

Other word classes rarely feature on the key word list; there are few numerals (*ena* 'one', *dve* 'two') and adverbs (*tukaj/tukajle* 'here', *najprej* 'first of all',[3] *značilno* 'typical').

Other noticeable words on the key word list are conjunction *torej* 'so' (found at the top of the ley word list), conjunction or pronoun *zakaj* 'why', preposition *med* 'among' (commonly used to introduce comparisons) and the filler *eem* 'um' (indicating pauses between long explanations).

3.2 Media-informative discourse

Few verbs feature on the key word list of media-informative discourse; those that do usually act as discourse markers (*po/glejte* 'look'), constitute similar pragmatic ex-

[3] It is not always possible to find an appropriate English translation with the same part-of-speech category as the one of the Slovene word.

pressions (*(moram) reči* 'I have to say'; *(kar se pa mene/tega) tiče* 'concerning this'), or perform discourse politeness strategies (*(če/a) dovolite* 'if you allow me'). Others express discourse acts such as agreement (*strinjam* 'I agree'), wish (*želel* 'I wish') and asking for opinion (*(kaj) pravite* 'what do you say'). Some forms of auxiliary verb 'be' are also on the list (past and future plural and conditional).

The largest group of key words is nouns; however, many are connected to the discourse topic (e.g. *vlada* 'government', *plače* 'salaries', *podjetja* 'companies'). Non-topical nouns are: *gospod* 'mister', *gospa* 'Mrs (missis)' and *gospoda/-om* 'sir', used by speakers to address each other, and *(tvoje/svoje/moje/vaše) mnenje* '(your/my) opinion' and *stališče* 'standpoint' for asking and expressing someone's opinion. Also, time appears to be quite an important element: *(tem/tistem/vsakem/zadnjem) trenutku* '(in this/each/last) moment' and *(v zadnjih/prihodnjih) letih* '(in the last/next) years' feature on the key word list. There are other nouns on the list such as *vprašanje* 'question', *dejstvo* 'fact', *možnost* 'possibility' etc.

A number of adverbs are found among the key words, such as *predvsem/zlasti* 'especially', *zelo* 'very', *veliko* 'a lot', *popolnoma* 'completely' etc. Some of these are employed as a sort of pragmatic expression (especially *pravzaprav* 'actually', *seveda* 'of course', *nenazadnje* 'after all'). The adverb *naprej* 'on' is common as part of the general extender *in tako naprej* 'and so on', whereas *potrebno/treba* 'need to' expresses necessity.

Adjectives are not very common on this key word list; those featuring express that something is *pomembno* 'important', that the speaker is convinced about something (*(sem) prepričan* 'I'm sure') etc. Numerals are also rare: we found only *tisoč* 'thousand' and *milijon* 'million', often used to express an amount of money or to indicate a year.

Typical pronouns are demonstrative (*tem/tega/teh* 'this'), possessive (*svoje* 'mine'), or indefinite (*vseh* 'all', *nekako* 'somehow' – the latter usually used as a marker of approximation: *je ta uporaba nekako dvajsetodstotna* 'is this usage somewhere around twenty percent').[4]

Many conjunctions and prepositions were detected as key words: conjunctions *in* 'and', *ki* 'that', *kot* 'as', *vendar* 'but' etc., and prepositions *v* 'in', *za* 'for', *o* 'about', *z* 'with' etc. Also, two particles were detected (*tudi* 'also' and *skratka* 'anyway'), as well as the interjection *hvala* 'thank you'. Last but not least, the filler *eee* 'um' was the first on the key word list of media-informative discourse.

3.3 Media-entertainment discourse

Many of the key words in the media-entertainment discourse are nouns. While some are topic-bound (e.g. *sezona* 'season', *mesto* 'city', *tekma* 'game' etc.), most are connected to the most common acts on the radio/TV program: announcing what will follow (*vreme* 'weather', *novice* 'news'), announcing music (*hiti* 'hits'), time (*ura* 'hour', *minut* 'minute'), telephone numbers for calls (*telefoni* 'telephones'), prize competitions (*(nagradna) igra* '(prize) competition), name of the show (e.g. *poslušate oddajo* 'you're

[4] The translation of the example phrase deviates from the direct translation of the word.

listening to the show’), name of the speakers (*(pred) mikrofonom* ‘at the microphone’) or radio station (*poslušate radio Siti* ‘you’re listening to radio City’), and asking for *aplavz* ‘applause’ on the TV. Particularly typical for media-entertainment discourse, especially on the radio, are greetings. Key nouns *jutro* ‘morning’*, dan* ‘day’*, večer* ‘evening’*, pozdrav* ‘greeting’ are mainly found in phrases *dobro jutro* ‘good morning’*, dober večer* ‘good evening’*, lep pozdrav* ‘greetings’ etc. To the same group belong interjections *čao* ‘ciao’ and *adijo* ‘bye’. The adjective *dragi* ‘dear’ is usually used in the phrases *dragi poslušalci/gledalci* ‘dear listeners/viewers’.

Jutro ‘morning’, as well as key adverbs *danes* ‘today’ and *zjutraj* ‘in the morning’, refer to the time. Adverbs are otherwise not very common among key words of this discourse type, e.g. *res* ‘really’ and *malo* ‘a bit’ were found on the list.

Verbs are not typical, with only a few featuring on the key word list. *Upamo* ‘we hope’ and *želimo* ‘we wish’ are among them, usually used in a positive context, e.g. *čudovit dan želimo* ‘we wish you a wonderful day’. *Dajte/dajmo* ‘let’s give’ is often used to encourage an applause (*dajte en aplavz* ‘let’s give an applause’), to form an imperative (colloquial usage, e.g.: *dejte se osredotočit* ‘let’s concentrate’), or for pragmatic uses, e.g. *ma dejte dejte Bruno* ‘oh come on come on Bruno’. *Vidiš* ‘you see’ is usually a discourse marker (e.g. *to je nevarno vidiš* ‘this is dangerous you see’). *Si* ‘yourself’ is either part of a verb phrase such as *vzeti si* ‘to take yourself’, or an auxiliary in singular, indicating the informal relationships between discourse participants (*si morda pogrešala* ‘did you perhaps miss’).

Several interjections also feature on the key word list; besides the aforementioned *čao* ‘ciao’ and *adijo* ‘bye’, we also found *bravo* ‘bravo’*, ooo, uuu,* and *ho*. These indicate that vocal expression of emotions is common; however, when used by professional speakers, emotions are most likely part of a performance. The interjection *hvala* ‘thank you’ is also on the key word list, as well as *no* ‘well’ and *evo* ‘you see’; both usually used as discourse markers.

Key conjunctions are *toda* ‘but’*, kajti* ‘for’*, torej* ‘so’ and *in* ‘and’. Key prepositions are *čez* ‘over’ and *pred* ‘in front of’ (often in the phrase *pred mikronofom* ‘in front of the microphone’ and to announce time: *pred eno uro* ‘an hour ago’). There is only one pronoun on the key word list, namely *nami* ‘us’ (*kdo je z nami* ‘who’s joined us’).

3.4 Official discourse

The key word list of the official discourse contains very few nouns, and those found are mostly topic-bound (e.g. *stopinj* ‘degrees’*, podjetje* ‘company’). Only the noun (used in a phrase) *(v) bistvu* ‘actually’ is topic-free, and its use is rather pragmatic, as sort of filler: *ne to v bistvu jst urejam* ‘no actually I arrange this’.

Some verbs feature on the key word list, but not many: *imate* ‘you have’*, razumem* ‘I understand’*, zdi* ‘it appears’ are among them; the verb (phrase) *(saj) pravim* ‘(as) I say’, and the verb *mislim* ‘I mean’*,* which is often used as a discourse marker (*ne mislim ni ni zdej moj problem* ‘no I mean this is not my problem’).

The largest word group among key words on this list is adverbs. Common are standard and colloquial forms of ‘then’ (*potem* and *pol*), in usages such as: *jah pol pa na-*

jboljš tko a ne ‘well then this is the best right?’. Adverb *tako* ‘so’ (or ‘more or less’) is also on the list, which, aside from its standard usages, can be used as: (1) a sort of backchannel, (2) a marker of vague language (e.g. *mene zanima tko mal na splošno* ‘I’m interested in this more or less in general’), or (3) part of the phrase *tako da* ‘so that’, often for the speaker to indicate continuity of the utterance (e.g. *letno poročilo pa še čakamo en delček tko da* ‘the year report we’re still waiting so that’). Adverb *tu* ‘here’ is not necessarily used to indicate a place, as deictic, but can be also used in situations such as this: *tk da tu bm reku* ‘so at this point I’ll say’; similarly to the adverb *tule* ‘here’ (pronounced as *tle*). *Zdaj* ‘now’ and *okej* ‘okay’ (and the corresponding *(v) redu* ‘all right’) are also on the key word list and are usually used as discourse markers.

There are many particles and interjections among the key words, often functioning as discourse markers; for example *ja* ‘yeah/yes’, *(a) ne* ‘y’know’, and *eem* ‘um’. Other particles are *pač* ‘indeed’, *mogoče* ‘maybe’, *recimo* ‘let’s say’. The greeting interjection *na svidenje* ‘good bye’ is also on the list. Conjunctions are *ker* ‘because’, *če* ‘if’, *saj* ‘as’ (in the phrase *saj pravim* ‘as I say’) and *pa* (which, besides its standard functions as ‘but’, is also a colloquial substitution for *in* ‘and’; e.g. *vlomijo pa kej odnesejo* ‘they break in and take something’).

Key pronoun *vam/vi* ‘you’ reveals that speakers often use more formal second person plural than less formal singular when addressing each other. However, *vi* ‘you’ is also often explicated, which is not acceptable in standard Slovene, e.g. *kolko rabite vi kapacitete* ‘how much capacity do you need’ (standard Slovene would be *koliko rabite kapacitete*).

3.5 Private discourse

There are almost no nouns among the key words of private discourse. Those found are marked by their colloquial origin: *fora* ‘joke’, *folk* ‘folk’ and *cajt* ‘time’, or form a phrase typical of the discourse: *(ni) problema* ‘(no) problem’, *(ni) panike* ‘no panic’, *(v) redu* ‘(all) right’. The same is true for adjectives: on the list are *fajn* ‘fine’, *ful* ‘full’ (phrases *ful dobro* ‘very well’, *ful fajn* ‘very fine’), *kul* ‘cool’, *(v) glavnem* ‘anyway’.

Verbs are more common; one large group comprises different forms of the verbs ‘go’ and ‘come’: *šla/šel/šli, iti, greš/grem, prišel, hodi*. In addition, more than one form of some other verbs feature among the key words: *imela/imel/imaš* ‘to have’; *rekel/rekla* ‘said’ (often used when telling a story, describing an event, or something that happened in the past); *videl/videla* ‘seen’, *glej* ‘look’ (usually a discourse marker: *lej kričat ti ni treba* ‘look you don’t have to scream’); *veš/vem* ‘know’ (forming phrases such as: *a veš* ‘do you know’, *ne vem* ‘I don’t know’, *vem ja* ‘yes I know’, *veš da* ‘no way’, *veš kaj* ‘you know what’). *Mogel/mogla* ‘could’ are often used instead of the standard *moral* ‘have to’ (*tako da si mogo čakat* ‘so that you had to wait’ – standard would be *tako da si moral čakat*), and *moreš* ‘have to’ is also on the list. *Daj* ‘give’ and *čakaj* ‘wait’ are often employed for pragmatic usages such as: *dej nehaj* ‘come on stop’ and *a čaki samo na eni strani je ne* ‘oh wait it’s only on one side’. Many forms of auxiliary *biti* ‘be’ are among the key word verbs, prevailingly in the singular form, indicating informal relationships between speakers.

Adverbs are quite common among key words in private speech. Many are usually used to refer to a place, as deictics (*tam/tja* ‘there’, *dol* ‘down’, *ven* ‘out’, *notri/noter* ‘in’), or time (*zmeraj* ‘always’). Some are considered colloquial words: *pol* ‘then’, *glih* ‘just’, *kao* ‘as’ and *ziher* ‘sure’, or are used in colloquial speech in a way that would be deemed inappropriate in more formal language: *gor* ‘up’ (*al ti pa lisice gor dene* ‘or he puts you handcuffs on’), *čisto* ‘totally’, *skoz* ‘all the time’ (*pa kaj je skoz gor na Fejsbuku* ‘is he on Facebook all the time?’), *drugače* ‘otherwise’ (*ma ne drugače je bil včeraj zelo lep dan* ‘well no otherwise it was a very nice day yesterday’), *res* ‘really’ (phrases *saj res* ‘of course’, *a res* ‘really?’), *itak* ‘sure’ (e.g. *ja itak* ‘well sure’). *Menda* ‘I guess’ also features on the list, usually expressing doubt.

Many personal pronouns figure among typical key words in private discourse: *jaz* ‘I’, *ti* ‘you’, *mene, meni* and *mi* ‘(to) me’, *mu* and *ga* ‘(to) him’, *oni* ‘they’, *ona* ‘she’, *on* ‘he’, and *ono, one,* and *onega* ‘it’ etc. This may be because speakers often talk personally about themselves or others. Furthermore, personal pronouns are often explicated (e.g. speaker says *jz sn rekla* ‘I said’); this usage is perceived less appropriate in standard Slovene. Third person personal pronouns in this discourse type are also used instead of the standard *tisti* ‘that’, e.g. *hodil po uni poti* ‘he walked on that road’. In addition, question pronouns *kaj* ‘what’ (used in phrases such as *veš kaj* ‘you know what’, *kaj praviš* ‘what you say’, *ali kaj* ‘or what’) and *koliko* ‘how much’ are among the key words; as well as demonstrative pronoun *tisto* ‘that’ and indefinite *vse* ‘all’. *Nič* ‘nothing’ and *ene* ‘about’ also feature on the list, the latter predominantly as a marker of approximation, e.g. *zdej se mi zdi da so ene štiri gor* ‘I think there are about four up there’.

Another very important group of key words in private discourse is interjections – *aja* ‘oh’, *ma* ‘well’, *ej* ‘ei’, *he* ’he’, *eh* ‘ah’, *joj* ‘ou’, *fak* ‘fuck’, *marija* ‘dear me’, *pizda* ‘fuck’, *kurba* ‘whore’ are on the list, mostly used for pragmatic functions (*ma ti puhlej* ‘well you look’) or in expressive speech (*o fak hudo carsko* ‘oh fuck totally cool’). Conjunctions are *pa* ‘but/and’, *saj* ‘but’ and *ko* ‘when’.

3.6 Discussion

When inspecting the concordances of each key word on our lists, we identified various reasons for why a particular word appears. Many words, especially nouns, appear on the key word lists because each of the five discourse types has its own specific and recurring setting and actions, e.g. classes are marked by teacher/lecturer vs. pupils/audience setting. Similarly, several words, mostly topic-bound nouns and verbs, feature on the lists because certain discourse types represent specific topics, discussed at length or repetitively (e.g. in classes), or deal with typical and recurring topics (e.g. in media-informative discourse).

Some words feature because of differences in the level of formality between different discourse types. The majority of such words (e.g. colloquial words or expressions) appear in private discourse. We also noticed that there may be differences between the five discourse types in terms of how much text is content-bound or content-free (metadiscourse). In non-formal types, there may be much more content-free text, therefore implying content-free use of words, such as discourse markers and other pragmatic expressions.

Several words appear on the key word lists as a result of being part of a commonly used phrase; though particularly common in private discourse, such words feature to some extent in all discourse types.

4 SPEECH DISCOURSES: DIFFERENCES AND SIMILARITIES

It is evident that the presence or absence of a particular key word group can be a good indicator of a type of discourse, or discourses (see Table 2). For example, nouns feature almost exclusively as key words in media discourses and classes discourse, with the usage of topic-free nouns being particularly typical of both media discourses. Personal pronouns are typical key words only in official and private discourses, whereas adjectives are typical key words only in media-informative and private discourses. As top key words, prepositions are limited to media discourses, and fillers to classes and media-informative discourses. Key word groups found as typical of only one discourse are particles and numerals.

Typical key words of three different discourses are pronouns, interjections, and conjunctions; although it is worth noting that the three discourses differ for each word group. With regards to interjections and conjunctions, and also prepositions, it is interesting that they are used to the same degree (i.e. the word groups contain a similar number of key words) across the discourses in which they occur.

The most widely typical word groups, found as key in all five discourses, are verbs and adverbs. It is interesting that key verbs are particularly typical of classes and private discourses, two quite different discourses in terms of the level of formality. The typicality of key adverbs is slightly more varied, but increases in less informal discourses.

	Classes	Media - informative	Media - entertainment	Official	Private
Topic-related key nouns	++	+++	+	+	
Topic-free key nouns		++	+++		
Key verbs	+++	+	+	+	+++
Key adjectives		+			+
Key adverbs	+	++	+	+++	++
Key numerals	+				
Key pronouns (without personal pronouns)	++	+		+	
Key personal pronouns				+	++
Key interjections			++	++	++
Key particles				++	
Key conjunctions		+	+		+
Key prepositions		+	+		
Key fillers	+	+			

Table 2: Keyword-based characteristics of the five spoken discourses.

+++ very typical
++ typical
+ found, but not very typical

The overall comparison of key words of the five spoken discourses, shown in Table 2, also reveals some similarities between discourses. Media-informative and media-entertainment discourses are similar in key word distribution of nouns, verbs, conjunctions and prepositions. This may be expected, given that both discourses use the same channels of communication (TV and radio). A certain level of similarity is found between official discourse and private discourse, especially in key word representation of adverbs, personal pronouns and interjections. Official discourse also shows similarities with media discourses, especially media-entertainment, in terms of key word presence of nouns (topic-related) and verbs. Similarly, private discourse, albeit quite different from many other discourses, shares some characteristics with media-entertainment discourse. Classes discourse displays characteristics of media discourses on one hand (e.g. key word representation of nouns, adverbs, pronouns and fillers), and of private discourse on the other (key word representation of verbs).

5 CONCLUSION

We employed the key word extraction method to identify statistically significant lexical items. We found that the presence or absence of a particular word class in the key word list can be a good indicator of a type of spoken discourse, or discourses. When the key word analysis is combined with manual analysis of concordances, it also provides valuable information about the characteristics of a particular type of discourse. This method of identifying discourses by word class of key words has several applications, for example, for corpus-building where texts could be automatically classified into a particular discourse on the basis of their lexical characteristics. To obtain clearer criteria for distinguishing between discourses, we plan to conduct a comprehensive study of all key words, as well as investigating key word phrases (i.e. multiword key words) and key key words (i.e. key words occurring in many different texts rather than in only one or a few).

References

Biber, Douglas/Stig Johansson/Geoffrey Leech/Susan Conrad/Edward Finegan (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas/Susan Conrad/Randi Reppen/Pat Byrd/Marie Helt (2002) "Speaking and Writing in the University: A Multidimensional Comparison." *TESOL Quarterly* 36/1, 9–48.

Biber, Douglas (2006) *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins.

Duguid, Alison (2010) "Newspaper discourse informalisation: a diachronic comparison from keywords." *Corpora* 2/10, 109–138.

Johnson, Sally/Astrid Esslin (2006) "Language in the news: Some reflections on keyword analysis using Wordsmith Tools and the BNC." *Leeds Working Papers in Linguistics* 11, 1–13.

Krajnc, Mira (2005) *Besedilne značilnosti javne govorjene besede: Na gradivu sej mariborskega Mestnega sveta*. Maribor: Slavistično društvo.

Kranjc, Simona (1999) *Razvoj govora predšolskih otrok*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Jeon, Jieun/Jae-Woong Choe (2009) "A Key Word Analysis of English Intensifying Adverbs in Male and Female Speech in ICE-GB." In: O. Kwong (ed), *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1.* Hong Kong: City University of Hong Kong, 210–219.

Lee, Joseph (2009) "Size matters: an exploratory comparison of small- and large-class university lecture introductions." *English for Specific Purposes* 28/1, 42–57.

Louwerse, Max/Scott Crossley/Patrick Jeuniaux (2008) "What if? Conditionals in educational registers. " *Linguistics and Education* 19/1, 56–69.

Nesi, Hilary (2001) "A corpus based analysis of academic lectures across disciplines. " In: J. Cotterill/A. Ife (eds) *Language Across Boundaries*. London: Continuum Press, 201–218.

Scott, Mike (1997) PC Analysis of Key Words – and Key Key Words. *System*, 25/1, 1–13.

Scott, Mike (1998) “Focusing on the Text and Its Key Words.” In: C. Stephens (ed), *TALC 98 Proceedings*. Oxford: Humanities Computing Unit, Oxford University, 152–164.

Scott, Mike (2000) “Focusing on the Text and Its Key Words.” In: L. Burnard/A. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt: Peter Lang, 103–122.

Scott, Mike (2001a) “Mapping Key Words to Problem and Solution” in M. Scott/G. Thompson (eds), *Patterns of Text: in honour of Michael Hoey*. Amsterdam/Philadelphia: Benjamins, 109–127.

Scott, Mike (2001b) Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In: M. Ghadessy/A. Henry/R. L. Roseberry (eds), *Small corpus studies and ELT: theory and practice*. Amsterdam/Philadelphia: Benjamins, 47– 67.

Scott, Mike (2002) “Picturing the key words of a very large corpus and their lexical upshots – or getting at the Guardian’s view of the world.” In: B. Kettemann/G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, 43–50.

Scott, Mike (2008) *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Scott, Mike/Chris Tribble (2006) *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins.

Smolej, Mojca (2007) Eliptične strukture in njihove metabesedilne vloge. *Jezik in slovsto* 52/3–4, 67–78.

Toolan, Michael (2004) “Values are Descriptions; or, from Literature to Linguistics and back again by way of Keywords.” *Belgian Journal of English Language and Literatures*, 11–30.

Tribble, Chris (2000) Genres, keywords, teaching: towards a pedagogic account of the language of project proposals In: L. Burnard/A. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt: Peter Lang, 75–90.

Verdonik, Darinka (2007) *Jezikovni elementi spontanosti v pogovoru: Diskurzni označevalci in popravljanja*. Maribor: Slavistično društvo Maribor.

Verdonik, Darinka (2010) Vpliv komunikacijskih žanrov na rabo diskurznih označevalcev. In: Š. Vintar (ed), *Slovenske korpusne raziskave.* Ljubljana: Znanstvena založba Filozofske fakultete, 88–108.

Verdonik, Darinka/Andrej ŽGANK/Agnes Pisanski Peterlin (2008) The impact of context on discourse marker use in two conversational genres. *Discourse studies* 10/6, 759–775.

Verdonik, Darinka/Ana Zwitter Vitez (2011) *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Volk, Jana (2007) Italijanske jezikovne prvine v spontanem govoru v Slovenski Istri. *Annales, Series historia et sociologia* 17/1, 153–166.

Zuljan Kumar, Danila (2007) *Narečni diskurz: diskurzivna analiza briških pogovorov*. Ljubljana: Založba ZRC.

Žgank, Andrej/Tomaž Rotovnik/Darinka Verdonik/Zdravko Kačič (2004) "Baza Broadcast News za slovenski jezik (bnsi) in sistem za razpoznavanje tekočega govora." In: T. Erjavec, J. Žganec Gros (eds), *Informacijska družba is'2004: Jezikovne tehnologije.* Ljubljana: Jožef Stefan Institute, 94–98.

Žgank, Andrej/Tomaž Rotovnik/Matej Grašič/Marko Kos/Damjan Vlaj/ Zdravko Kačič (2006) "Slovenska govorna baza parlamentarnih razprav za avtomatsko razpoznavanje govora." In: T. Erjavec, J. Žganec Gros (eds), *Informacijska družba is'2006: Jezikovne tehnologije.* Ljubljana: Jožef Stefan Institute, 115–118.

Žibert, Janez/France Mihelič (2004) "Development, evaluation and automatic segmentation of Slovenian Broadcast News Speech Database." In: T. Erjavec, J. Žganec Gros (eds), *Informacijska družba is'2004: Jezikovne tehnologije.* Ljubljana: Jožef Stefan Institute, 94–97.

Abstract
KEY WORD ANALYSIS OF DISCOURSES IN SLOVENE SPEECH: DIFFERENCES AND SIMILARITIES

One of the aspects of speech that remains under-researched is the internal variety of speech, i.e. the differences and similarities between different types of speech. The paper aims to contribute to filling this gap in research by making a comparison between different discourses of Slovene spontaneous speech, focusing on the use of vocabulary. The key word analysis (Scott 1997), conducted on a million-word corpus of spoken Slovene, was used to identify lexical items and groups of lexical items typical of a particular spoken discourse, or common to different types of spoken discourse. The results indicate that the presence or absence of a particular word class in the key word list can be a good indicator of a type of spoken discourse, or discourses.

Key words: corpus analysis, media discourse, private discourse, official discourse, spoken language.

Povzetek
ANALIZA KLJUČNIH BESED V DISKURZIH SLOVENSKEGA GOVORA: RAZLIKE IN PODOBNOSTI

Ena od značilnosti govorjenega jezika, ki ostaja še dokaj neraziskana, je njegova notranja raznolikost, torej razlike in podobnosti med različnimi vrstami govora. V tem prispevku se lotevamo te problematike s primerjavo različnih diskurzov spontanega slovenskega govora, pri čemer se osredotočamo na uporabo besedišča. Leksikalne enote, tipične za določen diskurz ali skupne različnim diskurzom, smo identificirali z analizo ključnih besed (Scott 1997), opravljeno na milijonskem korpusu govorjene slovenščine. Rezultati kažejo, da je prisotnost ali odsotnost besed določene besedne vrste na seznamu ključnih besed lahko dober kriterij za določanje vrste govorjenega diskurza.

Ključne besede: korpusna analiza, medijski diskurz, zasebni diskurz, nezasebni diskurz, govorjeni jezik.