Chapter 6

Cohesion through the lens of EPTIC-SI: Sentence-initial connectors in interpreted, translated and non-mediated Slovene

Tamara Mikolič Južnič[a] & Agnes Pisanski Peterlin[a]
[a]University of Ljubljana

Due to a lack of appropriate resources, few studies are devoted to comparing linguistic characteristics across different modes of production (such as speech and writing). This paper focuses on contrasting the use of sentence-initial connectors in mediated spoken and written Slovene and non-mediated spoken and written Slovene, by comparing EPTIC-SI, two monolingual reference corpora of Slovene, GOS for spoken and KRES for written discourse, and a subsection of a comparable Slovene corpus of parliamentary discourse, siParl. The EPTIC corpus and its subcorpus for Slovene, EPTIC-SI, are intermodal compilations of European Parliament speeches, their verbatim reports, interpretation transcripts and verbatim reports translations. This structure allows for direct comparison of the same content in different modes of production; however, the current size and monolithic genre of the corpus would make generalizations unreliable. For this reason, reference corpora of spoken and written discourse were used to complement the EPTIC corpus. The results show notable differences between the two modes of production, and at the same time reveal other influencing factors, such as genre and mediation.

1 Introduction

Traditional linguistic research on Slovene has focused above all on the standard written variety of the language, which means that there is much less data available on other varieties. This can present a challenge because, due to historical

Tamara Mikolič Južnič & Agnes Pisanski Peterlin. 2022. Cohesion through the lens of EPTIC-SI: Sentence-initial connectors in interpreted, translated and non-mediated Slovene. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Mediated discourse at the European Parliament: Empirical investigations*, 161–190. Berlin: Language Science Press.
DOI: 10.5281/zenodo.6977048

circumstances, there is a considerable gap between spoken Slovene and the standard written variety in terms of phonology, grammar and discourse. However, in recent decades, there has been increasing interest in compiling various kinds of corpora for Slovene to allow researchers an insight into different types of actual language use. At present, the majority of corpus resources available for Slovene focus on single modes of production, i.e., written texts (e.g., Gigafida,[1] KRES[2]), web discourse (e.g., Janes[3]) and spoken discourse (e.g., GOS[4]). While the compilation of new corpora has fostered a number of recent studies on non-standard and spoken Slovene (for instance, Fišer et al. 2020; Verdonik 2015), there has been less research interest in comparing the linguistic characteristics of Slovene across different language varieties or modes of production. One of the reasons for this may be that such comparisons are often difficult to carry out because the complex differences in content, genre, length, context, participants, etc. make direct contrasting of different types of materials challenging.

The present study attempts to address this gap. EPTIC-SI, the Slovene component of the EPTIC corpus, is used as a common platform for comparing the two modalities. A key advantage of EPTIC-SI is that it contains a spoken (interpreting) and a written (translation) version of the same content, which allows a direct comparison using a novel approach. Thus, data from EPTIC-SI can help us to shed light on how the written and spoken modalities of Slovene follow distinct discourse patterns. At present, a downside of EPTIC-SI is that it is a small corpus, further limited by the fact that it contains a single, monolithic discourse genre. As a result, it is difficult, if not impossible to generalize any findings based solely on its analysis. This means that complementing EPTIC-SI research with additional data from larger corpora helps increase the reliability and validity of the results.

This paper thus focuses on spoken and written varieties of mediated and non-mediated Slovene by comparing EPTIC-SI, two monolingual reference corpora of Slovene, as well as a Slovene corpus of parliamentary debates, siParl.[5] Specifically, we investigate variation in the use of sentence-initial[6] connectors, which constitute an important class of cohesive devices. We hypothesise that:

[1] http://www.gigafida.net/

[2] http://www.korpus-kres.net/

[3] http://nl.ijs.si/janes/o-projektu/korpus-janes/

[4] http://www.korpus-gos.net/

[5] https://www.clarin.si/noske/run.cgi/corp_info?corpname=siparl20&struct_attr_stats=1

[6] The term *sentence-initial* is used in this paper to refer to both written and spoken discourse, although *utterance-initial* or, in the case of a dialogue, *turn-initial* would be the appropriate terms for spoken discourse. A single term is used to simplify the comparison because the transcription conventions of the EPTIC-SI describe utterances as sentences.

- There is a difference between the use of sentence-initial connectors in interpreting and translation in the EPTIC-SI corpus.
- The use of sentence-initial connectors in interpreted Slovene in EPTIC-SI is similar to their use in spoken Slovene.
- The use of sentence-initial connectors in translated texts in EPTIC-SI is similar to their use in written Slovene.

The article is structured as follows: in §2, the compilation and the main characteristics of the new, Slovene component of EPTIC, EPTIC-SI, are presented. §3 is dedicated to a brief description of sentence-initial connectors to shed light on the topic under investigation. §4 presents a complete overview of the corpora and methods used. The results are presented and discussed in §5, followed by a brief conclusion in §6.

2 EPTIC-SI

EPTIC-SI is the Slovene component of the multilingual, parallel intermodal corpus known as EPTIC, or the European Parliament Translation and Interpreting Corpus,[7] comprising speeches delivered at the EU Parliament, their interpretations and translations (see Bernardini et al. 2016 for a more detailed description). The ongoing EPTIC project was first developed at the University of Bologna in collaboration with several other universities. As of 2020, the EPTIC corpus includes English, Italian, French, Slovene and Polish texts. With its intermodal and multilingual design, the EPTIC corpus fosters a range of different research perspectives, involving interpreting and translation and different types of comparisons of the different combinations of subcorpora. In addition, EPTIC allows the juxtaposition of interpretations and translations of the same content, facilitating a unique perspective on the differences between the two related yet divergent processes of interlingual communication.

At present, the Slovene language component of the EPTIC corpus, EPTIC-SI, is a collection of EU Parliament speeches, interpreted and translated into Slovene. Preselected speeches originally delivered in English on 17 January 2011 were used as source texts; the preselected speeches are the same speeches that are used in other parts of the EPTIC corpus. EPTIC-SI was compiled by a project team from the University of Ljubljana (UL) and UL MA-level students, consisting of Tamara

[7] https://corpora.dipintra.it/eptic/

Mikolič Južnič, Lia Lampe, Ana Podobnik, Polona Polc, Anina Stopinšek, Tamara Šiljak and Agnes Pisanski Peterlin. The preparation included the transcription of 64 selected speeches interpreted in Slovene and the preparation of metadata of both the transcripts and corresponding verbatim[8] reports translated into Slovene. In the narrow sense, EPTIC-SI thus consists of two subcorpora: 64 transcriptions of speeches interpreted into Slovene, and 64 written Slovene translations of English verbatim reports. For the purposes of comparison, two subcorpora comprising the corresponding source texts are also used (i.e. 64 transcriptions of original speeches delivered in English and 64 corresponding verbatim reports in English). The total number of texts in the four subcorpora is hence 256 and the total number of words is 76,445. All the components of EPTIC-SI have been aligned at sentence level and time-aligned with the video recordings in complete accordance with the EPTIC guidelines, and the data, along with the standardized metadata, is available from the main EPTIC webpage. In January 2020, the EPTIC-SI development began its second stage, with new materials being compiled to be added to the corpus by the end of the year.

3 Review of the literature

3.1 Sentence-initial connectors

Sentence-initial connectors have a significant role in text organization as they are used to link units of text; in fact, the importance of inter-sentential linking in establishing cohesion has long been recognized (see Halliday & Hasan 1976). There is no single agreed upon characterization or framework of connectors (see Halliday & Hasan 1976, van Dijk 1977, Fraser 1999), and there is considerable variation in terminology[9] (see, for instance, Crawford Camiciottoli 2010: 651). Nonetheless, it is generally agreed upon that they constitute a functional category (see Becher 2011: 30), which can be realized through a range of different linguistic elements (see Crawford Camiciottoli 2010: 650), including conjunctions (*and*), adverbs (*however*) and even phrases (*as a result*).

In recent years, a substantial body of corpus-based studies has provided novel insight into the function of connectors using authentic language. While some

[8] What is important to note is that, as Bernardini et al. (2016: 68) underline, even though they are called verbatim, "these reports are substantially edited" and may actually differ considerably from the transcripts, a fact, which must be taken into account when comparing the interpreted and the translated versions of the same speech (for a detailed account, see Bernardini et al. 2016: 62–70).

[9] Terms such as connectives, discourse markers, pragmatic markers and similar are used for this functional category by different authors.

of these studies focused on the complexities of corpus annotation (e.g., Rehbein et al. 2016, Crible 2017, 2018, 2017, 2020; Crible & Cuenca 2017; Crible & Pascual 2020), making an important contribution to identifying discourse-pragmatic phenomena in large collections of authentic texts, other studies explored the role of connectors in establishing cohesion in a text by investigating variation across languages, registers and discourse modes (Lapshinova-Koltunski & Kunz 2014; Kunz & Lapshinova-Koltunski 2014, 2015; Carrió-Pastor 2013).

Much research attention has focused on both intra-sentential as well as inter-sentential connectors, but the distinct discourse functions of sentence-initial connectors have also been highlighted and investigated (cf. van Dijk 1977; Moreno 1995; Dupont 2018). As Moreno (1995: 56) argues, sentence-initial connectors, functioning at the level of discourse, play a prominent role in the macrostructuring of the text. Several empirical studies on sentence-initial connectors, most notably Biber et al. (1999), have revealed a complex array of similarities and differences in terms of their use across spoken and written genres. The function of turn-initial connectors in spontaneous speech may, at first glance, appear quite distinct from the function of sentence-initial connectors in formal writing. However, as Dorgeloh (2004) shows, important parallels between the interactive discourse of spoken dialogue and writing can be established even for *and*, a connector that is far more frequent in speech than writing. Biber et al. (1999: 83–84) compare the frequency of use of selected "coordinators in sentence/turn-initial position" (*and, but, or* and *nor*) in conversation, fiction, news reportage and academic prose, revealing that they occur far more frequently in conversation than in any of the written registers, and that they occur least frequently in academic prose. Biber et al. (1999: 84) suggest that the somewhat more frequent use of coordinators in sentence-initial position in literature and news reportage may result from the fact that these two registers contain more dialogue. Biber et al. (1999: 83) also point out that there is "a well-known prescriptive reaction against beginning an orthographic sentence with a coordinator" (see also Dorgeloh 2004 and Bell 2007 for more information on the proscription against the sentence-initial *and* and *but* in English writing).

Finally, as EPTIC-SI comprises interpreted and translated discourse, the impact of mediation on connector use should also be considered, above all in terms of two mediation-related phenomena: transfer and explicitation. Transfer is the potential impact of source language conventions as reflected in the source texts on the target texts. Important cross-linguistic differences in the use of connectors have been identified for various pairs of languages in contrastive studies (e.g., Pit 2007; Lapshinova-Koltunski & Kunz 2014, Kunz & Lapshinova-Koltunski 2014, Balažic Bulc & Gorjanc 2015), including Slovene and English (Pisanski Peterlin

2015). The use of sentence-initial connectors in mediated discourse, including the texts in EPTIC-SI, may, in some cases, be influenced by transfer. Explicitation, i.e. the tendency of the target text to be more explicit than the source text, may result in an increased number of connectors in mediated texts. Musacchio & Palumbo (2010: 2) argue that "[c]onnectives are a good indicator of tendencies towards explicitation in translation as they can often be seen as optional elements". Furthermore, Gumul (2006) shows the importance and frequency of explicitation in the form of connectors in simultaneous interpreting, where they are largely a subconsciously added item in an automated mediation process. Based on the analysis of French-to-English and French-to-Dutch interpretations and translations of EU parliamentary speeches, Defrancq et al. (2015) confirm that interpreters add connective items for different reasons, including explicitation.

3.2 Hybrid speech-writing modes

Chafe & Tannen (1987: 383) underline that the difference between speech and writing began to receive research attention relatively late, as traditionally linguistics attempted to describe written language. The emergent focus on the differences between speech and writing has also contributed to a growing awareness that the relationship between the two modalities is not necessarily dichotomous. As Chafe & Tannen (1987: 391) argue, "there is no single feature or dimension that distinguishes all of speaking from all of writing". In this context, Wikström (2017: 30) highlights the contribution of the so-called "continuum" models, which "suggest that if particular registers such as everyday conversation and academic prose are taken as constituting poles of 'maximum' spokenness and writtenness respectively, most registers and genres of spoken and written discourse actually fall somewhere in between those poles as regards any given linguistic feature or discourse characteristic".

Outlining the fluid orality-literacy osmosis from a historical perspective, Soffer (2020: 930) touches upon the concept of secondary orality brought about by the advent of electronic media. He argues that"[i]n the electronic media age that followed print, texts are written to be read aloud". This type of blending of the two modalities is found in a range of hybrid genres, such as written-to-be-spoken discourse (e.g. pre-scripted speeches or television programmes), discourse spoken for transcription (e.g. medical dictation, intralingual live subtitling), digital Internet discourse (e.g. comments sections, tweeting) and mediated discourse (e.g. sight translation, interlingual subtitling). The blending is reflected in an array of linguistic features, ranging from lexical choice to syntactic complexity

(see Wikström 2017 for a detailed overview of the differences between the two modalities).

The intermodality of EPTIC provides a valuable insight into a genre that displays hybrid features of both spoken and written mode; as many of the speeches were pre-scripted, all were subsequently also transcribed as verbatim reports and underwent a substantial amount of editing.

4 Corpora and procedure

4.1 Corpora

In addition to the EPTIC-SI corpus, which comprises texts in both the spoken and written mode and is outlined in §2, two large reference corpora of Slovene comprised of written and spoken genres, as well as a comparable Slovene corpus of parliamentary debates were used in the study. These corpora were selected to enable a comparison between original and mediated texts in both modalities.

As described above, the EPTIC-SI corpus comprises EU Parliament speeches, and consists of transcripts of the Slovene interpretations of original English speeches and written translations of the English verbatim reports of the very same speeches.[10] While the original English speeches were not analysed in terms of connector use themselves, they were used to resolve any ambiguities about the function of a connector in the Slovene versions (only inter-sentential function was considered), as well as to shed light on whether the differences between interpretations and translations can be explained by the differences in the source texts. Table 1 summarises the statistical data for EPTIC-SI.

Table 1: Subcorpora and statistics for the EPTIC-SI corpus.

(Sub)Corpora	No. of words
EPTIC-SI English Spoken Sources (EPTIC SS)	21,561
EPTIC-SI English Verbatim Reports Sources (EPTIC VR)	20,552
EPTIC-SI Interpreting transcripts (EPTIC-SI Int)	16,143
EPTIC-SI Translated verbatim reports (EPTIC-SI Trans)	18,189
Total	76,445

[10] The speeches and the interpretations were thus produced before the verbatims and their translation.

The reference corpus of written Slovene is the KRES corpus, the 100-million word reference corpus sampled on the (much larger) Gigafida corpus. Since Gigafida, at present the biggest corpus of the Slovene language, is composed largely of newspaper and magazine articles, KRES was designed to be its balanced counterpart, in which various written genres are represented to reflect the actual ratio of different genres encountered in the everyday life by an average Slovene reader. The texts collected in the corpus were published between 1990 and 2011, and the samples of texts included in the corpus were chosen randomly (see Logar Berginc et al. 2012 for details). The taxonomy and statistics of the corpus are given in Table 2.

Table 2: Structure and statistics of the KRES corpus.

Subcorpora	**No. of words**
Printed publications	79,830,144
• Books	35,088,699
• Literature	17,030,038
• Non-fiction	18,058,661
• Periodicals	39,727,038
• Newspapers	19,919,327
• Magazines	19,807,912
Miscellaneous	5,014,206
Internet	20,001,001
• News portals	8,000,131
• Companies and institutions	12,000,870
Total	99,831,145

As can be seen from Table 2, KRES consists of 6 subcorpora: Literature, Newspapers, Magazines, Internet, Non-fiction (mainly specialized texts) and Miscellaneous (Misc.). The vast majority of texts are written in standard Slovene, though some of the subcorpora may contain texts with elements of spoken language, displaying elements of hybridity (for instance, the Literature subcorpus in some of the dialogues or the comments which are part of the Internet subcorpus). For the present study, all the subcorpora of KRES were analysed, as they represent a range of relevant genres. Further refinement of genre/source selection within each subcorpus would have been useful, but the online concordancer for KRES does not enable for it.

The reference corpus of spoken Slovene used in the study is the GOS corpus (see Verdonik et al. 2013 for more detailed descriptions). GOS includes around 120 hours of speech, transcribed in two versions (pronunciation-based and standardized), which are linked to the corresponding audio files. Samples of spoken Slovene were collected from all the regions of Slovenia between 2004 and 2010. In total, it contains around 1 million words, and it is, to date, the only reference corpus of spoken Slovene. The structure and statistics of the corpus are presented in Table 3.

Table 3: Structure and statistics of the GOS corpus.

Subcorpora	**No. of words**
Public	583,666
• Informative and educational	353,144
• Television	104,030
• Radio	95,117
• Personal contact	153,997
• Entertainment	230,522
• Television	104,955
• Radio	125,567
Non-public	451,435
• Non-private	155,893
• Telephone	33,862
• Personal contact	122,031
• Private	295,542
• Telephone	69,012
• Personal contact	226,530
Total	1,035,101

Table 3 shows that GOS comprises 4 subcorpora, but for the purposes of the present analysis, only two were used: the Public informative and educational subcorpus (henceforth Info-Ed) and the Non-Public Private subcorpus (henceforth Private). The two subcorpora were chosen because they represent two very distinct types of spoken language. The Info-Ed subcorpus comprises fairly formal spoken discourse that has often been pre-scripted or pre-prepared to some extent. Specifically, public informative discourse covers media discourse (i.e. television and radio news), while public educational discourse encompasses lectures (e.g. in

secondary schools and universities). The Private subcorpus represents the other end of the spoken continuum as it comprises spontaneous speech from private contexts, that is spontaneous conversation among family, friends and similar. While this is quite distinct from the genre of EPTIC-SI, it provides a valuable insight into the range of differences in Slovene spoken discourse.

As none of the genres in the reference corpora are directly comparable to the genre of the texts in EPTIC-SI, a set of texts from a comparable corpus of parliamentary discourse in Slovene, siParl,[11] was also analysed. SiParl is a 200-million word corpus comprising transcriptions of parliamentary debates of the Slovene National Assembly (see Pančur & Erjavec 2020 for details). SiParl includes different types of debates, such as regular sessions, urgent sessions, sessions of individual working bodies of the assembly, etc., with texts spanning three decades (1999–2018). During this period, Slovene society underwent a profound transition which may also be reflected in discourse characteristics. A small, relatively homogenous subsection of the corpus was carefully selected for a close comparison with EPTIC-SI. The 283,908-word subsection was limited to the genre of public presentation of opinions (henceforth Opinions), which is comparable to the genre of EPTIC-SI, and to the year 2011, also corresponding to the timeframe of EPTIC-SI. In making the selection, comparability was prioritised over size, with the restricted size of the subsection making manual analysis feasible.

4.2 Procedure

The criteria used to define sentence-initial connectors in this study were both formal (sentence initial position) as well as functional (discourse cohesive function). Halliday & Hasan (1976) identify four main types of conjunctive cohesion: additive, adversative, causal and temporal; in the present study, our analysis is limited to the first three categories, i.e. additive, adversative and causal.

For the purposes of corpus analysis, a list of 7 Slovene connectors was drafted for each of the three categories (see Appendix A). These lists were prepared in three steps. As relatively little data are available for Slovene on the linguistic items that can function as connectors and may appear in sentence-initial position, the first versions of the lists were compiled using several different sources. These included Toporišič's (2004: 646–652) list of intra-sentential coordinate conjunctions, Pisanski Peterlin's (2015) study of sentence-initial adversative connectors, Balažic Bulc & Gorjanc's (2015) study of the position of connectors and Hirci & Mikolič Južnič's (2014) study of causal connectors. The initial lists were further

[11] https://www.clarin.si/repository/xmlui/handle/11356/1236

expanded in the second step using the Slovene thesaurus function of Microsoft Word. The last step involved editing the list to retain only those connectors that unambiguously occur in intra-sentential function when used in the initial position. This was done because the size of the KRES corpus makes it impossible to manually examine all the results.

The searches were carried out automatically by means of the web concordancer for KRES, NoSketch Engine for GOS and siParl, and AntConc (Anthony 2020) for EPTIC-SI. The frequency counts were normalized to their rate of occurrence per 1000 words. For KRES, siParl and EPTIC-SI, where standard punctuation was used, determining the beginning of the sentence was not problematic. In GOS, double slashes marking the end of an utterance or a turn were used to identify utterance-initial or turn-initial connectors which were considered to be the equivalents of sentence-initial connectors in spoken discourse (see Dorgeloh 2004, for arguments supporting the comparability of sentence-initial connector use in speech and writing).

Next, all the selected sentence-initial connectors identified in EPTIC-SI, siParl and GOS were examined manually to remove any false results, i.e. cases in which the items from the search list had other functions. Such cases were extremely rare for additive and adversative connectors (only one such case was found in EPTIC-SI, with a total of 7 in siParl and 8 in GOS), and fairly rare for causal connectors (only 6 such cases were found in EPTIC-SI and a total of 88 in siParl and 142 in GOS).[12] For KRES, manual cleaning was not feasible because of the corpus size (100 million words) and the total number of concordances found (123,165). As a result, the figures for KRES are unrevised. However, if we assume that the percentage of false results is at least similar (and probably lower) to that in GOS, then the figures in KRES for causal connectors, the category where false results were the most common, probably contain somewhere around 3.65% false results.

The results for the different subcorpora were compared in terms of their overall frequencies, their frequencies for the different types of sentence-initial connectors and the frequencies of the individual connectors.

Finally, the results of the two subcorpora of EPTIC-SI were compared using NoSketch Engine available from the EPTIC website, where the parallel aligned versions are available, to establish the differences and similarities between the interpreted and translated versions. The corresponding transcriptions of the original English speeches and the English verbatims were also consulted when necessary as described in §4.1.

[12] The notable difference in size between the Slovene part of EPTIC-SI on the one hand, and siParl and GOS on the other, must be taken into consideration when interpreting these figures.

5 Results and discussion

The normalized quantitative results of the analysis of all the subcorpora are presented in Figure 1 below. The results are first presented as a total figure for each corpus and then separately by subcorpora.

The ratios of the three categories of sentence-initial connectors – additive, adversative and causal – are given for the individual subcorpora in Figure 2 below.

The results are compared and discussed in more detail in §5.1–5.3.

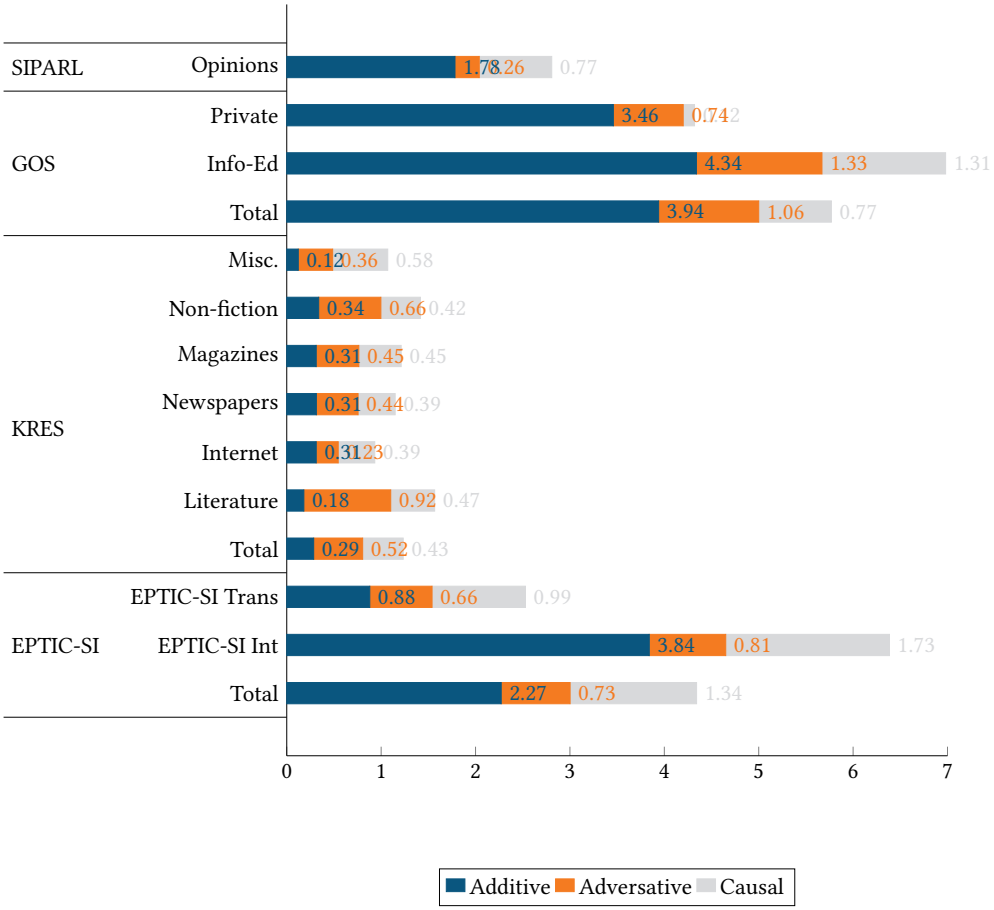

Figure 1: Occurrences of sentence-initial connectors in the analysed corpora.

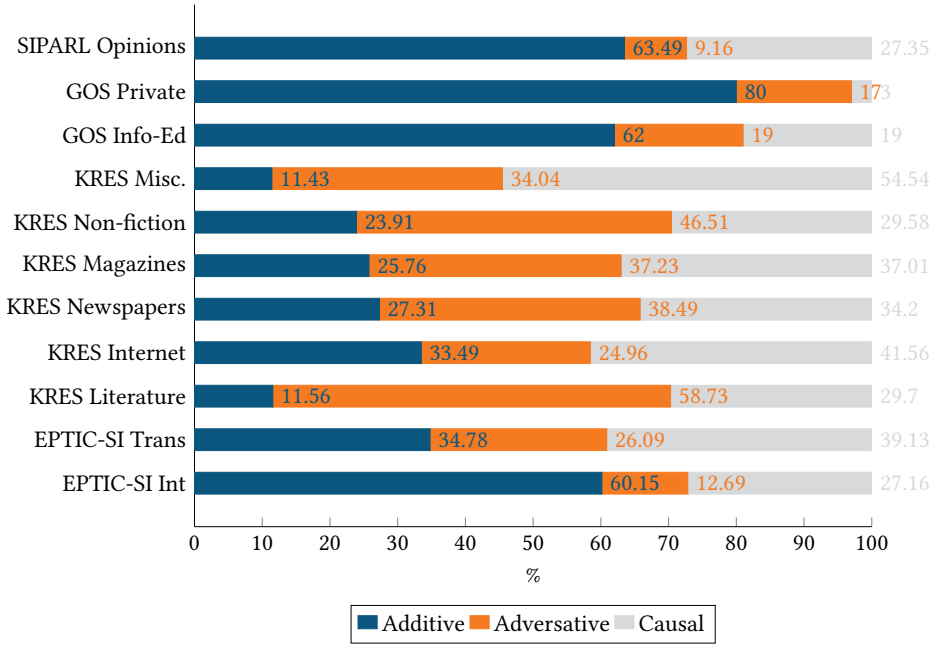

Figure 2: Ratios of the three types of sentence-initial connectors in individual subcorpora.

5.1 Sentence-initial connectors in EPTIC-SI Int and EPTIC-SI Trans

The first hypothesis examined in this paper is that there is a difference between the use of sentence-initial connectors in EPTIC-SI Int and EPTIC-SI Trans. The quantitative results of the corpus analysis are given in Table 4.

Table 4: Occurrences of sentence-initial connectors in the EPTIC-SI corpus

	Total EPTIC-SI		EPTIC-SI Int		EPTIC-SI Trans	
	Raw	/1k	Raw	/1k	Raw	/1k
Additive	78	2.27	62	3.84	16	0.88
Adversative	25	0.73	13	0.81	12	0.66
Causal	46	1.34	28	1.73	18	0.99
Total	149	4.34	103	6.38	46	2.53

The comparison of the interpreting and translation subcorpora of EPTIC-SI reveals a substantial difference in the frequency of use of sentence-initial connectors between the two subcorpora with the ratio being approximately 2.5:1, which confirms the first hypothesis. A juxtaposition of the three categories of connectors reveals that this difference is largely due to additive connectors, which occur four times as frequently in EPTIC-SI Int as in EPTIC-SI Trans. The difference in frequency is far less marked for adversative connectors that are used with almost the same frequency in both subcorpora. Finally, causal connectors are used almost twice as frequently in interpreting as in translation.

A more detailed focus on additive connectors shows that the marked difference is due to the use of a single connector, the sentence-initial *in* [and], which accounts for as many as 52 of the 62 additive connectors occurring in EPTIC-SI Int; only three other connectors, *poleg tega* [in addition] occurring 7 times, *prav tako* [additionally] occurring twice and *hkrati* [simlutanously] occurring once, are found in EPTIC-SI Int. In EPTIC-SI Trans, the most frequent additive connector is *poleg tega* [in addition] occurring in 8 cases, but other additive connectors are used rarely: *in* [and] in three instances, *obenem* [at the same time] twice, *prav tako* [additionally] twice and *ob tem* [at that] once. The preference for some of these connectors is closely linked to the register, as some connectors are very formal and associated with standard written texts, while others are more often used in sentence-initial position in informal contexts. However, as Dorgeloh (2004) argues, parallels between the discourse functions in speech and writing can be observed even in the case of sentence-initial *and*, which is far more frequent speech than writing.

When the results for the two subcorpora of EPTIC-SI are compared directly using the aligned versions on the EPTIC webpage, only three cases can be identified where there are matching additive connectors in both subcorpora in corresponding passages. A detailed look at the individual examples reveals that there are several other instances of matching sentence-initial additive connectors that cannot be identified automatically for various reasons, such as the use of a filler, *ehm*, immediately preceding the additive connector, but formally occurring in sentence-initial position (2 such cases), or the use of a less common sentence-initial connector not on the list used in corpus search (as in example (1)). But in the majority of cases, the manual check confirms that there are no matching sentence-initial connectors. In some of these instances, an intra-sentential additive connector is used in the corresponding passage in the other corpus, as in example (2). In other cases, no corresponding cohesive device can be identified.

(1) a. EPTIC-SI Int: *Istočasno* pa je pomembno poudariti tudi, da je Evropska unija eden največjih trgov za tropski les.[13]
[*Simultaneously*, it is important to stress that the European Union is one of the biggest markets for tropical wood.]

b. EPTIC-SI Trans: *Obenem* je tudi zelo pomembno, da poudarimo, da je EU eden izmed največjih trgov tropskega lesa.
[*At the same time*, it is very important for us to stress that the EU is one of the biggest markets for tropical wood.]

(2) a. EPTIC-SI Int: Mislim, da je to tudi eden od pomembni-, gre le za enega od kazalnikov, ampak če pogledamo celoto, zagotovo lahko govorimo o spodbudnih dogodkih. *In* edini način, da podpremo takšen proces, je da delamo skupaj z njimi …
[I think that this is one of the importa-, it is one of the indicators, but if we look at the whole, we can certainly speak of encouraging events. *And* the only way for us to support such a process is to work together with them …]

b. EPTIC-SI Trans: Razumem, da je to samo en kazalnik, a na splošno so bile novice vzpodbudne, proces *pa* lahko izboljšamo samo, če bomo sodelovali.
[I understand that this is only one indicator, but in general there has been encouraging news, *and* we can only improve the process by collaboration.]

There seem to be two main, often interrelated reasons for these omissions. The first is the register, or more specifically, the degree of formality. As certain additive connectors, above all *in* [and], are associated with speech and informal discourse, and are rarely used in formal, edited writing, it is not surprising that there are considerable dissimilarities in this area between the two subcorpora (example (1) illustrates such a difference in formality). The second reason is linked to the English originals. It is important to bear in mind that the interpretations and the translations are obtained using related but different source texts (see §2

[13] Throughout the text, the following markings are used for the examples from EPTIC-SI : a. for transcriptions of the Slovene interpretations of English speeches and b. for Slovene translations of the English verbatims. An English gloss, as literal as possible, is provided for all the Slovene examples in square brackets. Where necessary, c. for transcriptions of original English speeches and d. for English verbatims are added. In the examples, the relevant connectors have been highlighted in italics by the authors.

and Bernardini et al. 2016: 68): in spite of their name, the verbatim reports are heavily edited and diverge from the transcriptions of the speeches in terms of register and wording. As there is a strong proscription against using sentence-initial *and* in English (see Biber et al. 1999, Dorgeloh 2004, Bell 2007), it is not surprising that this is one of the features in which the source transcriptions and the verbatims in English differ greatly. Example (3) illustrates the difference between the two English versions, as well as the difference between interpreting and translation.

(3) a. EPTIC-SI Int: *In* še to za konec. Zelo hvaležna sem, da sem danes lahko predstavljala Evropsko komisijo pri tej točki. Podpredsednici Redingovi bom sporočila vse, kar ste povedali, tudi nekatera zastavljena vprašanja, vprašanje poslanca, kjer se pričakuje odgovor …

[*And* to finish. I am very grateful that I have been able to represent the European Commission on this topic today. I will convey to Vice-President Reding all of what you have said, including some of the questions, the question raised by an MEP where an answer is expected …]

b. EPTIC-SI Trans: Podpredsednici Reding bom tudi prenesla vse, kar je bilo povedano nocoj, vključno z vprašanjem, ki ga je postavil eden izmed poslancev in pri katerem se pričakuje odgovor.

[I will convey to Vice-President Reding all of what has been said today, including the question posed by one of the MEPs where an answer is expected.]

c. EPTIC SS: *And* my fifth and final point is that I'm very grateful that I have been here on behalf of the Commission this evening. I will convey to Vice-President Reding the points that have been made, including a question that has been raised here by one of the MEPs that that an answer is expected.

d. EPTIC VR: Finally, I will convey to Vice-President Reding the points that have been made here this evening, including the question raised by one member in relation to which an answer is expected.

At first glance, the comparison of adversative connectors reveals surprising similarities between examples in EPTIC-SI Int and EPTIC-SI Trans: 12 instances of the adversative connector *vendar* [however] occur in each subcorpus; in addition, there is only a single instance of another adversative connector *po drugi*

strani [on the other hand] in the interpreting subcorpus. Nevertheless, a juxtaposition of the two sets of examples shows, somewhat unexpectedly, that there are only two matching expression of *vendar* [however] in the two subcorpora. An examination of the remaining instances of *vendar* [however] in both subcorpora reveals that, for most of them, markers signalling adversative relations can be found in the corresponding passages of the translations and interpreted speeches. However, these markers are not identified through corpus search for several reasons: a) they are not used in sentence-initial position, b) they are not typical adversative connectors and are therefore not on the list of sentence-initial connectors used in this study, c) they may express adversative relations, but when used in sentence-initial position, they typically do not function as adversative connectors and are therefore not on the list used in corpus search. In about one third of the cases, no corresponding adversative marker can be identified in the parallel subcorpus. As in the case of additive connections, this often occurs when there is already a discrepancy between the transcription of the original English speech and the English verbatim, as in example (4) below.

(4) a. EPTIC-SI Int: *Vendar* pa dolgoročen cilj humanitarne pomoči ni ehm to.
[*But* the long-term goal of the humanitarian aid is not ehm that.]

b. EPTIC-SI Trans: Humanitarna pomoč pa seveda ni pravi instrument, ki bi imel dolgoročen vpliv.
[Humanitarian aid of course is not the right instrument that would have a long-lasting impact].

c. EPTIC SS: Ehm *but*, for long-lasting impact, humanitarian aid of course is not the instrument.

d. EPTIC VR: Of course, for a long-lasting impact, humanitarian aid is not the right instrument.

The omission of *but* in the verbatim can very likely be attributed to the proscriptions against using sentence-initial *but* in writing in English (cf. Bell 2007: 183); as Bell (2007: 194) points out this proscription is far less strong than the proscription against sentence-initial *and*, but it nevertheless needs to be taken into account. While there are no such restrictions against using *vendar* [however] in initial position in written Slovene, the fact that the Slovene translation is based on the English verbatim necessarily means that some of the adversative connectors are not found in the translations.

As with the other two categories, there is relatively little variety in causal connectors. Only three such connectors occur in the interpreting subcorpus: zato

[therefore] in 16 cases, *torej* [thus] in 7 cases and *zaradi tega* [because of that] in 5 cases, with 28 cases all together. In the translation subcorpus, all 18 causal connectors are instances of *zato* [therefore].

A close comparison of the results of the two subcorpora reveals that there are five matching causal connectors occurring in corresponding passages in both interpretations and translations. In several other cases, markers of causal or resultative relations can be found in the corresponding passages, often in the form of a clause, as in example (5). It seems that this reflects the complexity of the cause-effect relation which, unlike the additive meaning, tends to be overtly expressed.

(5) a. EPTIC-SI Int: *To je tudi razlog,* zakaj predvidevamo finančno pomoč za izboljšanje trgovskih zmogljivosti…

[*This is also the reason* why we expect financial aid for enhancing trade capacity…]

b. EPTIC-SI Trans: *Zato* je tu tudi finančna pomoč, ki bo okrepila trgovinsko zmogljivost.

[*Therefore,* financial aid is available to enhance trade capacity.]

Another interesting observation concerns the question of sentence boundaries and the parallels between intra-sentential and inter-sentential expressions of causality. It is noteworthy that when it comes to causal connectors, there are several instances where sentence boundaries diverge considerably between the interpreted speeches and the corresponding translations. In such cases, a sentence-initial causal connector would have a corresponding intra-sentential cause-result connector, as in example (6).

(6) a. EPTIC-SI Int: Želim odkrit razgovor z vami, sicer bom …*Torej* ehm vi ste v glavnem govorili tudi v angleščini, zato bom tudi jaz govoril v angleščini. Rekli ste, da naj si pogledamo…

[I wish to speak openly with your, otherwise I will …*So* ehm you have been mainly speaking in English, so I will speak in English as well. You have said that we should take a look…]

b. EPTIC-SI Trans: Besedilo imam v portugalščini, vendar bom improviziral v angleščini, *saj* ste v delu svojega govora, ki je bil po mojem mnenju najpomembnejši, uporabili ravno ta jezik…

[My text is in Portuguese, but I will improvise in English, *since* you have used this language in the part of your speech that I consider to be the most important part…]

Finally, a single case of a sentence-initial causal connector in EPTIC-SI Trans and a corresponding passage in EPTIC-SI Int with a combination of a sentence-initial additive connector *in* [and] immediately followed by a causal connector was found through corpus search (see example (7)). Once again, this type of difference clearly illustrates the disparity between less formal, more loosely organized spoken discourse (the metadata confirms that the speech in question is an impromptu speech), and structured, edited, written text.

(7) a. EPTIC-SI Int: *In zato* je treba pozdraviti z vsem srcem takšen sporazum in upam, da se bo tudi izvajal, kajti če se ne bo izvajal, bo škoda papirja, na katerem je napisan.

[*And therefore* this agreement should be welcomed wholeheartedly and I hope that it will be implemented, because if it isn't, it will not be worth the paper it is written on.]

b. EPTIC-SI Trans: *Zato* je ta sporazum treba pozdraviti odprtih rok in upam, da se bo tudi izvajal, kajti če se ne bo, potem ne bo vreden papirja, na katerem je napisan.

[*Therefore*, this agreement should be welcomed enthusiastically, and I hope that it will be implemented because if it is not, it will not be worth the paper it is written on.]

5.2 Sentence-initial connectors in interpreted and spoken Slovene

The second hypothesis tested was that the use of sentence-initial connectors in EPTIC-SI Int is similar to their use in spoken Slovene in GOS and siParl. The quantitative results are given in Table 5 and Figure 3.

Table 5: Occurrences of sentence-initial connectors in GOS, siParl and EPTIC-SI Int

	GOS						EPTIC-SI		SIPARL	
	Info-Ed		Private		Total		EPTIC-SI Int		Opinions	
	Raw	/1k	Raw	/1k	Raw	/1k	Raw	/1k	Raw	/1k
Additive	1533	4.34	1023	3.46	2556	3.94	62	3.84	506	1.78
Adversative	469	1.33	220	0.74	689	1.06	13	0.81	73	0.26
Causal	462	1.31	35	0.12	497	0.77	28	1.73	218	0.77
Total	2464	6.98	1278	4.32	3742	5.77	103	6.38	797	2.81

Two subcorpora of GOS were used in the present study. A comparison of the frequency of sentence-initial connectors in EPTIC-SI Int and in GOS (Total) shows considerable similarities. The frequencies of sentence-initial connectors in the comparable texts, siParl Opinions (public presentation of opinions from 2011), on the other hand, are considerably lower compared to both EPTIC-SI Int, as well as GOS and its subcorpora. However, as Figure 1 shows, the frequencies in siParl Opinions are still much higher than in all the written subcorpora of KRES, but only marginally higher than in EPTIC-SI Trans.

A closer look at the ratios of the three types of connectors for written and spoken corpora in Figure 2 reveals a clearer distinction between speech and writing in terms of sentence-initial cohesive devices.

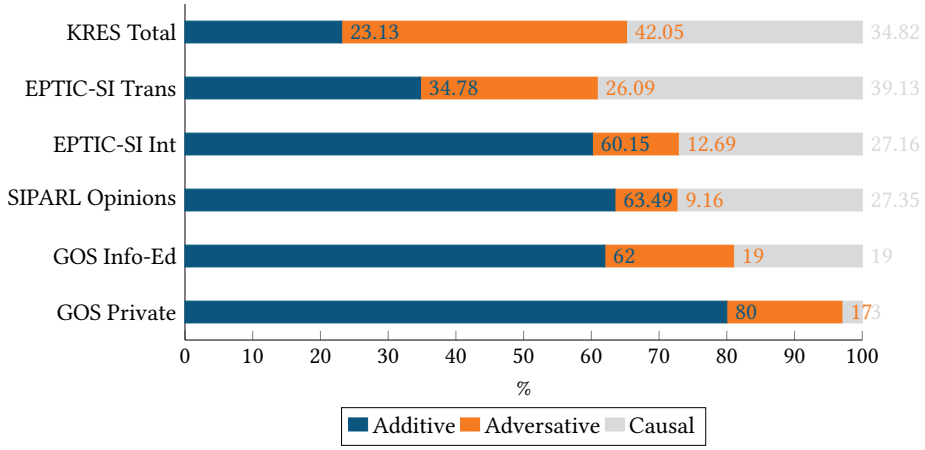

Figure 3: Ratios of the three types of sentence-initial connectors in spoken and written discourse.

Figure 3 reveals interesting distinctions between speech and writing. While sentence-initial additive connectors constitute the most frequently used category of connectors in all spoken subcorpora, this is not the case in the written texts of EPTIC-SI Trans and KRES, where causal and adversative connectors play a greater role in establishing inter-sentential cohesion. Moreover, the ratios in three of the spoken subcorpora, GOS Info-Ed, siParl Opinions and EPTIC-SI Int, are far more similar than in the fourth spoken subcorpus, GOS Private. This very likely reflects the fact that the Private subcorpus of GOS contains casual spontaneous conversation (see example (8)), an informal dialogical genre quite distinct from the content of EPTIC-SI Int. The discourse of the Info-Ed bears closer similarity to the genre of EPTIC-SI (see example (9)). The discourse of siParl Opinions (see example (10)) is, of course, most comparable to that of EPTIC-SI

Int, as both include structured, pre-prepared, formal and monological genre of parliamentary speeches. Nevertheless, the comparison with GOS Private offers an important insight into commonalities across a range of varieties of spoken discourse compared to written texts.

(8) GOS Private: //*in* kaj je narobe z njimi? / sandale // tiščijo me ona ma bl mičkano nogu ku jst // kdu? / *in* kaj pa če bi mi jih meni dala ? // [name:personal] // ja pomir si // [gap] // ja točnu tud ti pomir si // sej ne vem kire si mela

[//*and* what's wrong with them?/ sandals // they are too tight her feet are smaller than mine // who? / *and* what if you gave them me to me? // [name:personal] // well try them on // [gap] // well sure you try them on, too // I don't know which ones you had][14]

(9) GOS Info-Ed: // eee mislim da teh upov ni več eee vlada je na današnji seji sprejela sklepe s katerimi je dala soglasje za odprtje oziroma zaprtje štirih poglavij / *in hkrati* dala soglasje oziroma ne izdala soglasja za izd [gap] odprtje sedmih poglavij

[// erm I think that these hopes are long gone erm in today's cabinet meeting the government has passed agreements with which it gave its approval for the ope [gap] opening or closing of four chapters / *and at the same time* it gave its approval or denied its approval for the opening of seven chapters]

(10) Ustavite ga, tudi vi, gospod državni tožilec. Hkrati naj na koncu opozorim še na eno zadevo, ki se danes dogaja še vedno, mislim, da ni nobenih sprememb po prihodu novega generalnega državnega tožilca.

[Stop him, you too, Mr. Public Prosecutor. At the same time let me point out another matter that is still happening today, I believe there have been no changes after the arrival of a new general public prosecutor]

The relatively frequent use of causal connectors in the EPTIC-SI Int subcorpus might be explained by the fact that the genre of EU Parliament speeches is generally argumentative in nature and tends to use causal connectors as means of building arguments (cf. Didriksen & Gjesdal 2013). In siParl Opinions, the overall

[14]The annotations used in GOS include pauses, gaps, utterance beginnings/endings, etc. As noted in §4, utterance beginnings/endings and turn taking in dialogue are marked with double slashes, while pauses are marked using single slashes. However, it is essential to bear in mind that determining utterance boundaries is not as clear-cut as establishing sentence boundaries.

use of causal connectors is much lower than in EPTIC-SI Int; nevertheless, causal connectors constitute one quarter of sentence-initial connectors in both EPTIC-SI Int and siParl Opinions, underlying the argumentative character of parliamentary speeches. Interestingly, a comparatively high frequency of causal connectors occurs in the Info-Ed subcorpus of GOS, especially compared to the Private subcorpus, though the more diverse nature of the individual genres of Info-Ed (news reports, lectures), which may be more or less argumentative, probably accounts for the somewhat lower frequency of causal connectors than in EPTIC-SI Int. Moreover, connector use may be more frequent in interpreted texts due to explicitation and transfer (see §3), although a comparison with the corresponding original English texts, which is beyond the scope of the present paper, would be necessary to provide insight into translation-related phenomena.

The second hypothesis was thus partly confirmed: in terms of ratios, the results show a distinct cline with an overwhelming reliance on additive connectors in non-mediated spontaneous speech, and a more even distribution of the types of connectors in non-mediated writing. Although the frequencies of sentence-initial connectors also showed some degree of similarity among the spoken subcorpora, the tendencies are somewhat less homogenous.

5.3 Sentence-initial connectors in translated and written Slovene

The third hypothesis, that the use of sentence-initial connectors in EPTIC-SI Trans is similar to their frequency in written Slovene in KRES, is based on the assumption that the translated verbatim reports in EPTIC-SI follow the norms of written Slovene. As Table 6 shows, the quantitative results of our analysis support the third hypothesis only partially.

As noted in the Introduction, there is a substantial divergence between spoken and written genres in Slovene. The corpus data for KRES and GOS (see Figure 1) very much reflect this divergence between the two modalities, as the sentence-connectors analysed here occur far more frequently in spoken discourse. However, the comparison of the frequency of sentence-initial connectors in EPTIC-SI Trans and their overall frequency in KRES also shows a prominent difference: sentence-initial connectors are used twice as frequently in EPTIC-SI Trans as in KRES. A more detailed look at the categories of sentence-initial connectors shows that all are used less frequently in KRES, with the difference being particularly noticeable for additive and causal connectors.

The more frequent use of sentence-initial connectors in EPTIC-SI Trans may result from the hybrid nature of the source texts, i.e. the verbatim reports, which

Table 6: Occurrences of sentence-initial connectors in the KRES subcorpora and in EPTIC-SI Trans.

			Additive	Adversative	Causal	Total
KRES	**Literature**	**Raw**	3086	15673	7927	26686
		/1k	0.18	0.92	0.47	1.57
	Internet	**Raw**	6246	4655	7751	18652
		/1k	0.31	0.23	0.39	0.93
	Newspapers	**Raw**	6249	8805	7825	22879
		/1k	0.31	0.44	0.39	1.15
	Magazines	**Raw**	6185	8938	8886	24009
		/1k	0.31	0.45	0.45	1.21
	Non-fiction	**Raw**	6114	11895	7565	25574
		/1k	0.34	0.66	0.42	1.42
	Misc.	**Raw**	613	1826	2926	5365
		/1k	0.12	0.36	0.58	1.07
	Total KRES	**Raw**	28493	51792	42880	123165
		/1k	0.29	0.52	0.43	1.23
EPTIC-SI	**EPTIC-SI Trans**	**Raw**	16	12	18	46
		/1k	0.88	0.66	0.99	2.53

are based on speeches. As they are written to be delivered in the spoken mode, they share the characteristics of both written and spoken discourse.

A comparison with the individual subcorpora of KRES shows the same tendencies for the categories of additive and causal connectors and for the total number of connectors in each subcorpus. Adversative connectors, on the other hand, reveal a different picture: they are actually used more frequently in the Literature subcorpus of KRES (see example (10)) than in EPTIC-SI Trans, while their frequency is exactly the same in the Non-fiction subcorpus and in EPTIC-SI Trans. Of all the subcorpora of KRES, sentence-initial connectors are used most frequently in the Literature subcorpus, possibly reflecting the imitations of speech (dialogue) found in literature, as shown in example (11) (see also Biber et al. 1999: 84 for similar findings).

(11) Saj ni nič posebnega,« je priznal. »*Toda* nikamor drugam te ne morem odpeljati

[It's nothing special," he admitted. "*However*, I can't take you anywhere else]

(12) »Moj palček? *In* …kako veš, da sem mislila, da si pikapolonica?«
["My gnome? *And* … how do you know that I thought you were a ladybird?"]

To sum up, due to its hybrid nature, EPTIC-SI Trans exhibits a frequency of sentence-initial connectors that is quite different from the spoken genres analysed as well as from the written genres in KRES, albeit the results are somewhat closer to those of the KRES corpus, compared to spoken genres. However, in addition to the influence of genres outlined above, another potential reason for the relatively high frequency of sentence-initial connectors in the EPTIC-SI Trans corpus should be considered. As it contains mediated discourse, explicitation of cohesive links as well as transfer from the source texts may well have contributed to the fairly frequent use of sentence-initial connectors in EPTIC-SI Trans.

6 Conclusion

The aim of the present study was to contrast the use of sentence-initial connectors, an important category of cohesive devices, both in spoken and written Slovene as well as in mediated and non-mediated discourse. Using EPTIC-SI, two large reference corpora for Slovene and a subsection of a comparable Slovene corpus of parliamentary discourse, we have shown that patterns of use of sentence-initial connectors reflect important differences for both dimensions, modality and mediation, thus substantiating the potential of this type of corpus research. The expected difference between mediated spoken and written discourse in the first hypothesis was confirmed, but the second and third hypotheses were only partly confirmed. For spoken non-mediated and mediated discourse, the results show a greater complexity, as the similarities depend on the type of connector. The written mediated discourse of EPTIC-SI Trans appears to display hybrid characteristics of both spoken and written discourse.

The EPTIC corpus offers a unique perspective on different modes of interlingual mediation and the complexities of language use, as it provides the same content in two different modalities and multiple languages. For Slovene as a peripheral language, the contribution of EPTIC-SI is particularly valuable because it enables us to directly observe and reflect on the differences between the same content worded in speech and writing, opening a range of additional research paradigms. We believe that the present study corroborates the multidimensional investigation potential of EPTIC and EPTIC-SI, providing insight into the intricacies of language reality.

Finally, the specific characteristics of language identified in EPTIC-SI may also shed light on other important issues in future research. The varieties of languages evolving in EU contexts, shaped by a variety of factors, including language mediation, have already been recognized as distinct forms of language production for other languages, most notably English (see, for instance, Trebits 2009, whose study focuses on the use of conjunctive cohesion in EU documents in English). However, the specific features of Slovene as used in EU contexts have not yet received systematic research attention; in fact, there seems to be little research awareness of new patterns developing in administrative and public discourse in Slovene as a result of the language contact in EU institutions. It therefore seems that as EPTIC-SI is gradually expanded, also to include original Slovene speeches delivered at the EU Parliament, it will offer an invaluable resource for studying this emerging new variety of Slovene.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0218 and No. P6-0215).

Appendix A List of sentence-initial connectors used in the corpus search

- Additive connectors:
  - In
  - Hkrati
  - Obenem
  - Ob tem
  - Poleg tega
  - Prav tako
  - Podobno
- Adversative connectors:
  - Na drugi strani
  - Nasprotno
  - Po drugi strani

  - Toda
  - Vendar
  - Vendarle
  - V nasprotju
- Cause-result connectors:
  - Kot posledica
  - Posledično
  - Torej
  - Zaradi tega
  - Zategadelj
  - Zato
  - Zatorej

References

Anthony, Laurence. 2020. *AntConc*. http://www.laurenceanthony.net/software/antconc/ (28 July, 2020).

Balažic Bulc, Tatjana & Vojko Gorjanc. 2015. The position of connectors in Slovene and Croatian student academic writing: A corpus-based approach. In Sonja Starc, Carys Jones & Arianna Maiorani (eds.), *Meaning making in text: Multimodal and multilingual functional perspectives*, 51–71. London: Palgrave Macmillan UK. DOI: 10.1057/9781137477309_4.

Becher, Viktor. 2011. When and why do translators add connectives?: A corpus-based study. *Target* 23(1). 26–47. DOI: 10.1075/target.23.1.02bec.

Bell, David M. 2007. Sentence-initial *And* and *But* in academic writing. *Pragmatics* 17(2). 183–201. DOI: 10.1075/prag.17.2.01bel.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of written and spoken English*. Harlow: Longman.

Carrió-Pastor, María Luisa. 2013. A contrastive study of the variation of sentence connectors in academic English. *Journal of English for Academic Purposes* 12(3). 192–202. DOI: 10.1016/j.jeap.2013.04.002.

Chafe, Wallace & Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology* 16(1). 383–407. DOI: 10.1146/annurev.an.16.100187.002123.

Crawford Camiciottoli, Belinda. 2010. Discourse connectives in genres of financial disclosure: Earnings presentations vs. earnings releases. *Journal of Pragmatics* 42(3). 650–663. DOI: 10.1016/j.pragma.2009.07.007.

Crible, Ludivine. 2017. Discourse markers and (dis)fluency in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics* 22(2). 242–269. DOI: 10.1075/ijcl.22.2.04cri.

Crible, Ludivine. 2018. *Discourse markers and (dis)fluency: Forms and functions across languages and registers*. Amsterdam: John Benjamins. DOI: 10.1075/pbns.286.

Crible, Ludivine. 2020. Weak and strong discourse markers in speech, chat, and writing: Do signals compensate for ambiguity in explicit relations? *Discourse Processes* 57(9). 793–807. DOI: 10.1080/0163853X.2020.1786778.

Crible, Ludivine & Maria-Josep Cuenca. 2017. Discourse markers in speech: Distinctive features and corpus annotation. *Dialogue & Discourse* 8(2). 149–166. DOI: 10.5087/dad.2017.207.

Crible, Ludivine & Elena Pascual. 2020. Combinations of discourse markers with repairs and repetitions in English, French and Spanish. *Journal of Pragmatics* 156. 54–67. DOI: 10.1016/j.pragma.2019.05.002.

Defrancq, Bart, Koen Plevoets & Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo (ed.), *Yearbook of corpus linguistics and pragmatics 2015: Current approaches to discourse and translation studies*, 195–222. Cham: Springer. DOI: 10.1007/978-3-319-17948-3_9.

Didriksen, Anders Alvsåker & Anje Müller Gjesdal. 2013. On what is not said and who said it: Argumentative connectives in Nicolas Sarkozy's speeches to the European Parliament. In Kjersti Fløttum (ed.), *Speaking of Europe. Approaches to complexity in European political discourse*, 85–110. Amsterdam & Philadelphia: John Benjamins.

Dorgeloh, Heidrun. 2004. Conjunction in sentence and discourse: Sentence-initial *and* and discourse structure. *Journal of Pragmatics* 36(10). 1761–1779. DOI: 10.1016/j.pragma.2004.04.004.

Dupont, Maïté. 2018. Between lexis and discourse: A cross-register study of connectors of contrast. In Sebastian Hoffmann, Andrea Sand, Sabine Arndt-Lappe & Lisa Marie Dillmann (eds.), *Corpora and lexis*, 173–208. Leiden: Brill. DOI: 10.1163/9789004361133_008.

Fišer, Darja, Nikola Ljubešić & Tomaž Erjavec. 2020. The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation* 54(1). 223–246. DOI: 10.1007/s10579-018-9425-z.

Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics* 31(7). 931–952. DOI: 10.1016/S0378-2166(98)00101-5.

Gumul, Ewa. 2006. Explicitation in Simultaneous Interpreting: A strategy or a by-product of language mediation? *Across Languages and Cultures* 7(2). 171–190. DOI: 10.1556/Acr.7.2006.2.2.

Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London & New York: Routledge.

Hirci, Nataša & Tamara Mikolič Južnič. 2014. Korpusna raziskava rabe vzročnih in pojasnjevalnih povezovalcev v prevodih iz angleščine in italijanščine. In Agnes Pisanski Peterlin & Schlamberger Brezar, Tamara (eds.), *Prevodoslovno usmerjene kontrastivne študije*, 150–70. Ljubljana: Znanstvena založba Filozofske fakultete.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2014. Cohesive conjunctions in English and German: systemic contrasts and textual differences. In Lieven Vandelanotte, Kristin Davidse, Caroline Gentens & Ditte Kimps (eds.), *Recent advances in corpus linguistics*, 229–262. Amsterdam: Rodopi. DOI: 10 . 1163 / 9789401211130_012.

Kunz, Kerstin & Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14(1). 258–288. http://ojs.ub.gu.se/ojs/index.php/njes/article/view/3095 (5 February, 2021).

Lapshinova-Koltunski, Ekaterina & Kerstin Kunz. 2014. Conjunctions across languages, registers and modes: Semi-automatic extraction and annotation. In Ana Diaz-Negrillo & Francisco Javier Diaz-Pérez (eds.), *Specialisation and variation in language corpora*, 77–104. Bern: Peter Lang.

Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt & Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: Gradnja, vsebina, uporaba*. Ljubljana: Trojina, Zavod za uporabno slovenistiko.

Moreno, Ana I. 1995. Causal intersentential relations: A discourse as process view. *Estudios ingleses de la Universidad Complutense* 3. 55–68. https://digital.csic.es/handle/10261/12407 (9 February, 2021).

Musacchio, Maria Teresa & Giuseppe Palumbo. 2010. Following norms, taking risks: A study of the use of connectives in a corpus of translated economics articles in Italian. In Carmen Heine & Jan Engberg (eds.), *Reconceptualizing LSP: Online proceedings of the XVII European LSP symposium*, 1–11.

Pančur, Andrej & Tomaž Erjavec. 2020. The siParl corpus of Slovene parliamentary proceedings. In Darja Fišer, Maria Eskevich & Franciska de Jong (eds.), *Proceedings of the Second ParlaCLARIN Workshop*, 28–34. Marseille: European Language Resources Association. https://www.aclweb.org/anthology/2020.parlaclarin-1.6 (5 February, 2021).

Pisanski Peterlin, Agnes. 2015. Sentence-initial adversative connectives in Slovene-English translation of academic discourse: A corpus study. In Mojca Schlamberger Brezar, Limon, David & Gruntar Jermol, Ada (eds.), *Contrastive analysis in discourse studies and translation.* 68–82. Ljubljana: Znanstvena založba Filozofske fakultete.

Pit, Mirna. 2007. Cross-linguistic analyses of backward causal connectives in Dutch, German and French. *Languages in Contrast* 7(1). 53–82. DOI: 10.1075/lic.7.1.04pit.

Rehbein, Ines, Merel Scholman & Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 23–28. Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2016/summaries/457.html (5 February, 2021).

Soffer, Oren. 2020. From textual orality to oral textuality: The case of voice queries. *Convergence: The International Journal of Research into New Media Technologies* 26(4). 927–941. DOI: 10.1177/1354856519825773.

Toporišič, Jože. 2004. *Slovenska slovnica [Slovenian grammar].* Maribor: Založba Obzorja.

Trebits, Anna. 2009. Conjunctive cohesion in English language EU documents – A corpus-based analysis and its implications. *English for Specific Purposes* 28(3). 199–210. DOI: 10.1016/j.esp.2009.04.004.

van Dijk, Teun A. 1977. *Text and context explorations in the semantics and pragmatics of discourse.* New York: Longman.

Verdonik, Darinka. 2015. Internal variety in the use of Slovene general extenders in different spoken discourse settings. *International Journal of Corpus Linguistics* 20(4). 445–468. DOI: 10.1075/ijcl.20.4.02ver.

Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek & Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation* 47(4). 1031–1048. DOI: 10.1007/s10579-013-9216-5.

Wikström, Peter. 2017. *I tweet like I talk : Aspects of speech and writing on Twitter.* Karlstads: Karlstads universitet. (Doctoral dissertation). http : / / urn . kb . se / resolve?urn=urn:nbn:se:kau:diva-64752 (15 February, 2021).