

# The European Parliament as a discourse community: its role in comparable analyses of data drawn from parallel interpreting corpora

BART DEFRANCO

EQTIS, Ghent University (Belgium)

## Abstract

*The aim of this paper is to explore the concept of comparability in corpus-based Interpreting Studies and, more in particular, the risk of using comparable components of parallel interpreting corpora. Quite a number of studies based on such an approach have yielded inconclusive results, which could be due, it is argued, to the fact that the comparable components are drawn from plenary sessions of the European Parliament (EP), in which the groups generating the data share a working environment and could therefore influence one another.*

*To investigate the potential linguistic convergence that is likely to result from this mutual influence, both a theoretical and an empirical approach are taken. The theoretical approach seeks to determine whether EP interpreters and members of the EP could be analysed as constituting one single discourse community, according to the criteria put forward by Swales (1990). The empirical study of three discourse markers, based on data from a parallel corpus of EP interpreting and a comparable corpus of British parliamentary debates, aims to yield evidence of linguistic conversion.*

## Keywords

Simultaneous interpreting, European Parliament, discourse community, discourse markers.

## Introduction

The aim of this paper is to explore the concept of comparability in corpus-based Interpreting Studies. This concept is crucial in much corpus-based work on translation and interpreting. Its use in corpus-based research in interpreting may be problematic in some respects.

Comparable corpora are text collections that contain “components that are collected using the same sampling frame and similar balance and representativeness” (McEnery *et al.* 2006: 48). From a principled point of view, comparable corpora are bilingual or multilingual and are used in contrastive and comparative research. More recently, and especially in the domain of Translation and Interpreting Studies, it is accepted that comparable corpora consist of texts from different varieties of one language. These varieties can be geographical (such as in the ICE or *International Corpus of English*<sup>1</sup>), diachronic, register-based, etc. The comparable corpora used in Translation Studies are of the latter type and consist of texts in one language that are either translated or non-translated. Mona Baker’s call to move the study of translation away from the comparison of source and target texts and to consider the translated text in its own right as a register of the target language (Baker 1993) to be studied in comparison with other registers, marked the start of an extremely fruitful period of corpus research.

Yet, many comparable corpora used in the research that followed Baker’s call are actually by-products of parallel corpora. When texts and translations are compiled for one language pair in two directions, the corpus automatically indeed contains non-translated and translated texts in two languages. This is usually illustrated by means of the diagram in Figure 1, taken from Johansson *et al.* (1999-2002) representing the different components of the *English-Norwegian Parallel Corpus*.

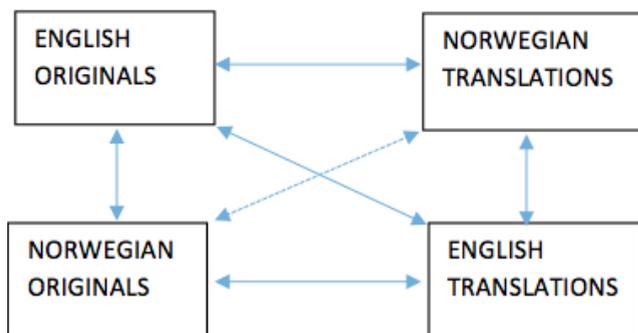


Figure 1. Structure of the *English-Norwegian Parallel Corpus*, from Johansson *et al.* 1999-2002:13

1 <<http://www.ucl.ac.uk/english-usage/projects/ice.htm>>.

In Figure 1, the horizontal solid arrows connect parallel sub-corpora; the vertical solid lines connect multilingual comparable sub-corpora and the diagonal solid line connects the monolingual comparable sub-corpora of translated and non-translated text. The configuration of sub-corpora can be expanded with other languages.

A considerable portion of the translation research carried out following Baker's proposals draws its data from this kind of configuration of sub-corpora<sup>2</sup>. In Interpreting Studies, its position is nearly hegemonic: an overwhelming majority of publications on translation universals in interpreting or even, more broadly, features of interpreted language are based on comparable corpora included in larger parallel datasets (Sandrelli/Bendazzoli 2005; Russo *et al.* 2006; Kajzer-Wietrzny 2012, 2015; Defrancq 2016; Defrancq *et al.* 2015; Bernardini *et al.* 2016; Ferraresi/Miličević 2017). Shlesinger (2009) and Shlesinger/Ordan (2012) are notable exceptions, but their data are experimental and can therefore not be considered particularly representative of the interpreting activity.

I will argue that such an approach, though inspired by reasonable practical considerations, may be problematic and particularly problematic in the case of Interpreting Studies. I will first discuss a number of biases in comparable research carried out on data drawn exclusively from a parallel corpus. I will focus, in particular, on the bias caused by potential mutual influence between groups whose linguistic patterns are recorded in the parallel corpus, as is the case of the European Parliament, where most of the interpreted corpus data come from (§1). Secondly, I will explore the theoretical basis for this mutual influence (§ 2). The next section (§ 3) will be devoted to the results of an empirical investigation of potential linguistic convergence. These results will be put into perspective (§ 4), after which the general conclusions of the research will be drawn (§ 5).

## 1. Comparability of components in a parallel corpus

Evidently, comparable corpora drawn from parallel corpora do not entirely meet one of the fundamental criteria comparable corpora are supposed to meet, namely the requirement that the sampling method be identical for the corpora to be compared. Admittedly, there are few comparable corpora based on identical sampling frames, as identical sampling usually comes at the cost of representativeness (Leech 2007). Cultural differences may make it difficult to find the same genres in different languages and the coverage of particular genres which is culturally appropriate in one language can be completely disproportionate in another.

However, problems run deeper in comparable datasets drawn from parallel corpora. First of all, there is one sampling criterion applied to the non-translated dataset that does not hold for the translated dataset: texts assumed to represent non-translated language are included on the condition that they were translated into the other language(s) of the corpus. This is obviously not the case of the

2 Interestingly, Mona Baker herself proceeded otherwise in Olohan/Baker (2000), a study on optional *that* carried out on a corpus of translations and a comparable set of data from the *British National Corpus* (BNC).

translated texts in the same language. In that sense, datasets drawn from parallel corpora can never be completely comparable.

In addition to this fundamental design issue, comparable datasets of interpreting face scholars with other important problems. I will not discuss the sizes of the corpora, which are obviously Lilliputian compared to what is mainstream in corpus linguistics and even corpus-based Translation Studies, nor the evident lack of balance in the sampling of speakers, as the number of speakers on the interpreting side is always considerably smaller than on the side of the source speakers. Neither of these problems is solvable without considerable investment. The issue I want to raise here is the context from which most of the data are drawn: the European Parliament (EP).

The only corpus data of interpreting including a comparable dimension<sup>3</sup> are indeed those included in what is sometimes called the “EPIC suite” (Bernardini *et al.* forthcoming), a series of corpora compiled at different universities (Bologna, Ghent, Poznan and Saarbrücken) and based on simultaneous interpretations carried out at the EP (hence EPIC: *European Parliament Interpreting Corpus*). The original EPIC was compiled in Bologna and contains more or less 200,000 tokens, source speeches included; the Poznan corpus (Kajzer-Wietrzny 2012) counts 250,000 tokens (without source speeches) and EPICG, compiled in Ghent currently boasts 220,000 tokens, source speeches included.

One could easily call into question the representativeness of interpreting at the EP. After all, due to the EP’s rules of procedure and limited speaking time allotted to the members of the EP (MEPs), interpreter turns are extremely short compared to what is common in other contexts and interpreters face high to very high speaker delivery rates. In addition, most MEPs read out written texts they prepare in advance, which requires interpreters to not only switch languages but also to switch registers. These contextual constraints are regularly brought up by researchers faced with evidence contradicting their hypotheses (Kajzer-Wietrzny 2012). The collected data are thus probably representative of EP interpreting only and generalisations about the interpreting activity – the ultimate purpose for which corpus research is carried out – are compromised.

However, these constraints do not raise doubts on the comparable nature itself of existing interpreting corpora. The real risk hides in another corner: in the corpora of the EPIC suite, non-interpreted and interpreted data are produced by speakers who not only share the same working environment while their speech is being recorded, i.e. the EP’s plenary room; they also listen a fair amount of time to one another through headphones. In those circumstances, it would be rather surprising if both groups did not influence each other’s linguistic output, in which case the so-called comparable corpora of speeches and interpretations are not independent data sets.

3 Parallel interpreting corpora without a comparable dimension have been developed at the University of Hamburg: CoSi (consecutive and simultaneous interpreting) and DiK (dialog interpreting in public service settings). CoSi covers interpretation between German and Portuguese, comprising 35,000 tokens, while DiK includes 165,000 tokens of German, Portuguese and Turkish, only part of which are interpreting data.

Strangely enough, the literature does not explicitly put forward independence of data sets as a criterion for compiling suitable comparable corpora, but it seems reasonable to avoid the confounding effect of dependence when conducting comparative or contrastive research on languages or language varieties. Its omission in the literature is probably merely due to the fact that in most cases dependence is not an issue, because authors or speakers represented in different corpus components are unlikely to have met physically. Admittedly, there is a very slim chance that translators into a particular language have met or read authors writing in the same language, but the scale of this exposure is incomparably smaller than the exposure in the EP.

The potential drawback of dependent data sets is evident: linguistic convergence, i.e. the levelling out of specific features of both types of output, making the whole search for features typical of interpreting ultimately pointless. This risk is not only theoretical: it is striking that many studies using comparable EP data fail to establish significant differences between linguistic patterns observed in MEPs' speeches and in EP interpretations. In one of the first comparable studies conducted on interpreting data drawn from the EP, Sandrelli/Bendazzoli (2005) observe contradictory tendencies both with regard to lexical density and with regard to lexical variation: no consistent pattern emerges as to which variety of Italian and English – non-interpreted or interpreted – presents higher levels of lexical density and variety than the other. Instead, lexical density in interpreting appears to be determined by the source language interpreters are working from, while differences in lexical variety seem to depend on the target language. In a recent study, Ferraresi/ Miličević (2017) compare Italian EP speeches and Italian EP interpretations with regard to the conventionality of phraseology. The many results of this detailed study almost invariably suggest that there is no significant difference in the conventionality of interpretations and speeches. The authors do not offer an explanation for this remarkable finding, focusing rather on differences between EP translations and EP interpretations, but it would not be unreasonable to blame linguistic convergence in the work place. Fair enough, some studies find significant differences between MEPs' speeches and EP interpretations on a number of parameters (Kajzer-Wietrzny 2012 notably), but generally the results are much less predictable than in comparisons between non-translated and translated sub-corpora.

Obviously, causality is hard to establish in these cases. It would require a lot more evidence of different types and a solid theoretical framework to conclude that linguistic convergence takes place between MEPs and EP interpreters. I will start with the latter in Section 2.

## 2. Theoretical explorations

Several theoretical and descriptive models are based on the idea that speakers influence each other's linguistic output. Communication accommodation theory (CAT) (Giles *et al.* 1991), for instance, holds that in conversation the properties of speakers' linguistic output tend to either converge or diverge, depending on the relationship that is being built among them. Pettel (2005), for example, observes

that convergent tendencies can be observed between speakers in the use of discourse markers. However, CAT is first and foremost a framework for the analysis of conversations, while oral communication at the European Parliament in contexts with interpreters is anything except conversation.

Other models based on concepts such as discourse communities (Swales 1990) or communities of practice (Wenger 1998) are more promising in this respect. They specifically apply to more institutionalised ways of communicating, recognising the importance of the professional environment in shaping the linguistic features of the discourse. However, the concept of communities of practice focuses on collective learning processes that occur in groups when members of the group engage with one another with regard to a shared interest or passion. Duflo (2014) convincingly shows that EU interpreters constitute a community of practice, but it would probably be an unacceptable stretch of the imagination to hold that MEPs and interpreters in the EP constitute one. Mutual engagement between the two groups is minimal and there is little to no collective learning effort. Some institutional initiatives are taken to improve MEPs' understanding of interpreting (Duflo 2014), but these probably do not count as instances of "mutual engagement", as they are only intended to show one group how to help the other group optimise its practice.

That leaves us with the discourse community. According to Swales (1990) a discourse community is a group of people whose shared linguistic repertoire is essentially determined by function (Swales 1990: 24). Swales puts forward six definitory features, among which the pursuance of common objectives by the community is the most important and also the one that discourse communities share with communities of practice. Other criteria include the presence and use of mechanisms of intercommunication, which seems to be a broader concept than Wenger's mutual engagement, the use of a participatory mechanism for information and feedback, a threshold level of members with discursual expertise, and, particularly interesting for our interests, specific genres and lexical repertoires.

There is little doubt that the European Parliament constitutes a discourse community. In her study on trends in the translation of the European Parliament's verbatim reports between Spanish and English, Calzada Pérez (2007) finds that EP membership meets the six criteria put forward by Swales (1990): the European Parliament pursues common public goals laid down by the Treaties, although members and parties can have conflicting interests; it possesses communicative and participatory mechanisms laid down in its Rules of Procedure; it possesses genres (plenary and commission meetings, verbatim reports,...) and a specific terminology, which it shares with other EU institutions. Calzada Pérez (2007) does not fully elaborate Swales' point about threshold levels of expert members, which I will raise below, suggesting the EP may be an atypical discourse community, because membership is not granted on the basis of expertise, but rather through universal suffrage. The two features that are of immediate relevance to our topic are the existence of specific genres and lexical repertoires in discourse communities, because they bear directly upon linguistic features that can be investigated on the basis of corpus data. Calzada Pérez, for instance, assumes that transitivity patterns, as observed in a corpus of verbatim reports, reflect linguistic choices typical of discourses developed in the EP.

However typical or atypical the EP discourse community may be, it is crucial to underline that in Calzada Pérez' study, the EP translators do not seem to be part of it. Their contribution to the communicative mechanisms in the EP is studied exclusively in its textual dimension, i.e. through translation shifts. The question that arises at this point is: is an alternative analysis of the discourse community at all possible? Is it theoretically possible to include EP translators in the EP discourse community? Does the same hold for interpreters? For, if our purpose is to account for specific linguistic patterns shared by MEPs and interpreters, it needs to be shown that interpreters belong to the discourse community which is presumed to generate these patterns.

It seems that the question can be answered affirmatively: translators and interpreters actively participate in the pursuance of the EP's goals. Given the EU's commitment to multilingualism and the linguistic rights MEPs draw from it, it would be nearly impossible for the EP to reach political decisions without interpreters or translators. Communicative mechanisms at the EP evidently involve interpreters and translators, as they ensure a crucial part of the communication. Their role is not merely technical in this respect. Communication also takes place between interpreters and translators on one side and the EP's administration on the other: documents are made available for preparation, questions can be asked regarding source texts (even though the many drafters are hard to identify, see Robinson 2014), etc. Admittedly, this communication is limited and highly institutionalised. Direct interaction between translators/interpreters and MEPs is probably close to non-existent, although interpreters do have the opportunity to communicate technical problems during meetings and are occasionally addressed by MEPs (Bartłomiejczyk 2017).

Participatory mechanisms are in place for feedback and improvement of overall performance. Besides the previously mentioned feedback on technical problems, DG INTE, the EP's interpreting directorate, offers MEPs documentation on how to successfully conduct meetings with interpreters and presentations about interpreting (Dufrou 2014). Interinstitutional guidelines are available for drafters of legislation, with a special focus on drafting for translation (European Communities 2003). As to the threshold level of expert members, translators and interpreters working at the EP can be considered experts, not so much regarding the actual content of legislation, which most MEPs are not experts in either, but regarding the functioning of the discourse community itself: its agenda-setting, meeting protocols, text cycle, etc. In terms of seniority, translators and interpreters probably even have an edge on MEPs, as MEP turnover is considerable: only 302 out of the 751 pre-election MEPs in 2014 were still in office by April 2017<sup>4</sup>. Interpreters and translators evidently master the terminology used at the EP. Translators in EU institutions, including the EP, are involved with the standardisation of terminology in their respective languages (Robinson 2014).

4 The list of the elected MEPs in the 2014 election can be found here: <[http://www.europarl.europa.eu/pdf/elections\\_\\_results/ElectedMEPs.pdf](http://www.europarl.europa.eu/pdf/elections__results/ElectedMEPs.pdf)>, as well as an indication as to whether the elected members have been re-elected or not. Incoming and outgoing MEPs since the 2014 elections are listed here: <<http://www.europarl.europa.eu/meps/en/incoming-outgoing.html>>.

It thus seems reasonable to conclude that the discourse community at the EP comprises interpreters and translators, besides MEPs. Interpreters and translators are admittedly peripheral as they do not contribute directly to the advancement of the community: unlike MEPs they do not have the ability or power to change the nature, the evolution or the prospects of the community. On the other hand, they do crucially contribute to the maintenance of the community by facilitating access to it for MEPs. Without the linguistic services they offer, MEPs would probably face more obstacles to fully participate in the community. That said, translators are probably slightly more peripheral than interpreters, as the latter share the working context, i.e. meeting rooms, of MEPs to a considerable extent, while the former are physically separated from the workings of MEPs.

The one aspect that was not discussed in assessing translators' and interpreters' involvement in the EP discourse community is genre. Swales (1990) views genres as linguistic and textual means by which the operations of the discourse community are articulated. New discourse communities need to establish genres and novices need to acquire the genres' properties to be considered members. Arguably, EP translators and interpreters are the most influential community members when it comes to the properties of genres in the EP. Of course, they do not determine the ways committee and plenary meetings are organised (Calzada Pérez 2007), but they do bear responsibility for the bulk of speech and text production at the EP: MEPs original speeches and texts only constitute a fraction of the EP's textual production. Most of the speech produced during committee or plenary meetings is interpreted speech and most of the circulated texts are translations. This implies that the linguistic properties of the EP genres are predominantly shaped by interpreters and translators. Therefore, even though interpreters' and translators' roles in the EP discourse community are peripheral regarding most of Swales' criteria, their impact on the development of discourse routines is important. This is especially the case for interpreters. In committee and plenary meetings, when it comes to their own mother tongue, MEPs probably spend more time listening to interpreters than to other MEPs they share their mother tongue with. Within these genres, their own practice is thus very likely to be influenced by interpreters' linguistic output. This brings us to a crucial point: in the shaping of oral genres and discourse patterns, it is most likely the interpreters who play a dominant role, not the MEPs.

Summing up, an analysis including interpreters and translators in the EP discourse community seems compatible with Swales' (1990) theoretical framework on discourse communities. The fact that they constitute peripheral groups in some respects does not fundamentally compromise the analysis, as discourse communities do not need to be homogeneous (Calzada Pérez 2007). The inclusion of interpreters in the EP discourse community provides a strong explanatory hypothesis for linguistic convergence patterns between interpreters and MEPs. Given the structure of the interaction between MEPs and interpreters, it furthermore seems plausible that interpreters play a determining role in shaping oral discourse at the EP.

### 3. Searching for evidence

In Section 1, it was suggested that the inconclusive results of some comparable studies of interpreting data could be ascribed to linguistic convergence between MEPs and EP interpreters. The framework of the discourse community provides a theoretical basis for such an analysis: discourse communities create genres and genres manifest themselves through particular linguistic features. However, the aforementioned studies focus on parameters such as lexical density and lexical variety, which do not seem to be directly connected to the concept of genre. That is why a different option will be taken here, i.e. the analysis of discourse marker use.

Within the framework applied here, the study of discourse markers is indeed especially relevant. Discourse markers have been shown to be genre-sensitive (Lam 2009; Aijmer 2013). Specific patterns of discourse marker use are believed to belong to the linguistic repertoire of discourse communities, to such an extent that these patterns can be felt to be “strange, incomprehensible or even absurd” (Lassen 2003: 158) outside the discourse community. Therefore, if the EP discourse community comprises both MEPs and EP interpreters, both groups will probably present converging patterns of discourse marker use. These patterns are expected to differ from patterns observed in groups that do not belong to the EP discourse community. In other words, different corpora and sub-corpora will have to be used: a sub-corpus of speeches held by MEPs, a sub-corpus of interpretations carried out by EP interpreters and a corpus of speeches collected from another context.

The EP data for this study were drawn from EPIC and EPICG, respectively compiled at the Universities of Bologna and Ghent. Only English data were used: English speeches delivered by British and Irish MEPs (EPIC) and interpreting performed in the English booth from French (EPICG), Spanish (EPIC) and Italian (EPIC). The other context used to compare the EP data with is the British Parliament. Data from the British Parliament were obtained through the BNC-BYU portal and probably offer the data that is most comparable to data from the European Parliament. It should be noted, however, that the BNC data are considerably older than the data from the EP. The former were based on parliamentary sessions held in 1994, whereas, depending on the corpus used, the EP data were transcribed from sessions held in 2004 (EPIC) and 2008 (EPICG). Another potential confounding factor is interaction: debates in the British Parliament tend to be much more interactive and less scripted than EP debates, where speakers are given a certain amount of time to basically hold a monologue most of them have prepared in written form. Table 1 shows the token count for each of the corpora.

	<b>Non-mediated BNC</b>	<b>Non-mediated EN EP</b>	<b>Interpreted FR&gt;EN</b>	<b>Interpreted ES&gt;EN</b>	<b>Interpreted IT&gt;EN</b>
<b># tokens</b>	96,239	45,523	37,566	13,905	7,202

Table 1. (Sub)-corpus sizes

For practical reasons, I will focus on 3 discourse markers: *well*, *now* and *so*. They are the most frequent discourse markers of English and therefore likely to occur reasonably frequently even in nano-corpora, such as the ones used here.

### 3.1. *Well*

*Well* is a usual suspect in corpus studies of English. Its use has been thoroughly investigated using different theoretical frameworks (Bakhtinian heteroglossia, Aijmer *et al.* 2006; relevance theory, Blakemore 2002) and different types of corpora, including written, spoken and learner corpora. Importantly, it has been established that *well* is register-sensitive (Aijmer 2013), making it a suitable item to investigate linguistic patterns potentially shaped by membership of a discourse community. Its many possible translations have been analysed in parallel corpora and its use or omission by interpreters has been noted in a number of studies (Hale 2004; Mason 2008; Blakemore/Gallai 2014; Defrancq 2016).

In a previous study on *well* (Defrancq 2016), carried out on EPIC and EPICG data, I was able to observe that despite the lack of specific triggers, interpreters use *well* quite frequently. Most occurrences (85%) of *well* lack a source trigger altogether. As Table 2 shows, its relative frequency in EP interpreting is approximately 0.5 occurrences per thousand words, slightly but not significantly lower than its frequency in MEPs' speeches (X-squared: 1,4618,  $p=0.52$ ). The figures are consistent across source languages (X-square: 0.85;  $p=0.66$ ; Fisher Exact:  $p=0.99$ ). In addition, with a few exceptions, the functions of *well* in interpreted speech were found to be identical to its functions in non-interpreted speech in the European Parliament.

Frequency	Non-mediated BNC S_Parliament	Non-mediated EN EP	Interpreted FR>EN	Interpreted ES>EN	Interpreted IT>EN	Total EP
#	76	25	15	8	4	52
/1000 w	<b>0.79</b>	<b>0.55</b>	0.40	0.58	0.56	<b>0.50</b>

Table 2. Frequencies of *well* in respective (sub-)corpora.

In addition to the data used in the previous study, Table 2 also shows the frequency of *well* in the parliamentary section of the BNC. It appears that the frequency of *well* in the BNC (S\_Parliament) is well above its frequency in EPIC and EPICG combined, the difference being statistically significant (X-squared = 6.62;  $p=0.01$ ). On the other hand, if only the non-mediated data from the BNC and EPIC/EPICG are compared (MEPs vs. British MPs), significance is not reached (X-squared = 2.51;  $p=0.11$ ). So MEPs and interpreters, as a group, present lower frequencies of discourse marker use, but, within that group, English-speaking MEPs resemble British MPs most.

With regard to function, *well* is most frequently used to introduce answers to questions and replies to statements in all sub-corpora. Due to the interaction-

al properties of the EP debates, where MEPs and invitees consecutively take the floor for a monologue that lasts 1 to 6 minutes, answers to questions are usually embedded in a rhetorical structure figuring both the question in a quoted format and the answer to it (example (1) from the EP source text sub-corpus). This structure is extremely rare in the British Parliament, where speakers may interact directly (example (2), representing British parliamentary discourse).

- (1) // ehm ... Mrs Oomen-Ruijten then asked me did we follow follow the Member States' lead in this // **well** the actual chronology of events is that on the twenty-third of January ehm the Commission put in place the ban in circumstances I have already outlined // [EPIC org-en 7]
- (2) [...] what was your view regarding the proposals that are being made? (SP:PS5PG) **Well**, we feel that there's (pause) it's obviously quite a lot of points [...] [BNC K77]

All other functions are represented in all sub-corpora: *well* is used to introduce quoted speech, to initiate the turn, to signal disalignment with other points of view expressed in previous discourse. Only in interpreting and in the British Parliamentary corpus is *well* also used for self-repairs and as a hesitation marker. The lack of such examples in MEPs' discourse is probably due to the highly scripted nature of the texts they present: they are prepared in advance in written form and read out during the plenary session. Finally, as mentioned in Defrancq (2016), *well* also occurs once at the start of an episode of incorrect interpretation. As mentioned before, in only 15% of the cases is *well* triggered by a source item. Interpreters seem to use it autonomously.

### 3.2. *Now*

The discourse marker *now* has also attracted a fair amount of attention in the literature. Broadly speaking, *now* used as a discourse marker has a topical function. It relates text segments to larger informational units: comparisons, lists, etc. (Schiffrin 1987), typically initiating new topics within larger structures. Register-sensitivity has been demonstrated on spoken and written academic corpora (Biber 2006). In Interpreting Studies, the spectacular omission rate of *now* in court interpreting was observed by Hale (2004).

Table 3 shows the frequencies of *now* in the respective corpora and sub-corpora used.

Frequency	Non-mediated BNC S_Parliament	Non-mediated EN EP	Interpreted FR>EN	Interpreted ES>EN	Interpreted IT>EN	Total EP
#	80	9	30	12	6	57
/1000 w	0.83	0.20	0.80	0.86	0.83	0.55

Table 3. Frequencies of *now* in respective (sub-)corpora.

The figures are remarkably different from those regarding the use of *well*. MEPs and EP interpreters present very different usage patterns ( $X^2 = 19.77$ ;  $p < 0.0001$ ), while the group as a whole, uses *now* to a degree that significantly differs from the British MPs ( $X^2 = 5.92$ ;  $p = 0.01$ ). However, this is exclusively due to the low frequency of *now* in MEPs' speeches, as interpreters and British MPs do not significantly differ ( $X^2 = 0.04$ ;  $p = 0.83$ ). Again, the use of *now* is consistent across source languages ( $X^2 = 0.05$ ;  $p = 0.97$ ).

In all sub-corpora, *now* functions predominantly as a "continuative" marker (Halliday 1994), marking shifting points between successive phases in the discourse (example (3) from the BNC), and topical shifts specifically, as in (4) drawn from the interpreting corpus.

(3) Now (pause) there seems to be a very close relationship between (pause) the European peoples party and the Conservative party, if only on that circumstantial basis. **Now** the European peoples party manifesto (pause) calls for a single currency [BNC JSF]

(4) the council is very happy with that // **now** as far as the timetable is concerned the position taken by the commissioners on the second of July arrived too late for us [EPICG \_\_25.09.08 \_\_paquetsocial \_\_roselynebachelotnarquin \_\_I \_\_en]

Other less frequent functions include the announcement of evaluative statements and, especially, statements in disalignment with other points of view, and acknowledgment of hearer's point of view. The range of functions *now* can have, appears to be broader in interpreting, where *now* is also found to signal turn uptake, the start of a question and as a hesitation marker, functions that are usually fulfilled by *well*. In only 33% of the occurrences of *now* in interpreting, can a source item be identified. Again, it is fair to conclude that interpreters use the marker fairly autonomously.

### 3.3. *So*

The uses of *so* have been extensively studied, many a scholar underlining the extreme polyfunctionality of the discourse marker (Buysse 2012). Corpus studies typically find that *so* is only found in a minority of cases to establish consequential relationship. Additional functions include textual organisation, such as signaling a return to the main argument of the text and various interpersonal uses, such as signaling turn management (Raymond 2004), other-orientedness (Bolden 2006), etc. Register-sensitivity has been shown by Lam (2009) on oral data. In previous work (Defrancq *et al.* 2015) on discourse marker use in simultaneous interpreting in the EP, we were able to conclude that the use of *so* by EP interpreters was not triggered by the source text in more than half of the occurrences. In most of these cases interpreters added *so* to explicitate an implicit discourse relationship they perceived in the source text. Some cases, on the other hand, were clearly motivated by an attempt to create an illusion of cohesion where the interpreters face difficulties.

Table 4 shows the frequencies of *so* in the EP, both for the English booth and the MEPs, in combination with its frequency in the BNC (S \_\_ Parliament).

Frequency	Non-mediated BNC S_Parliament	Non-mediated EN EP	Interpreted FR>EN	Interpreted ES>EN	Interpreted IT>EN	Total EP
#	113	38	100	21	23	182
/1000 w	1.17	0.83	2.66	1.51	3.19	2.45

Table 4. Frequencies of *so* in respective (sub-)corpora.

Again, MEPs and EP interpreters appear as different groups (X-squared = 38.6;  $p < 0.0001$ ), while the difference between MEPs and British MPs does not reach the significance threshold (X-squared = 3.35;  $p = 0.06$ ). As a group, MEPs and EP interpreters present patterns of use that differ from those observed in British MPs (X-squared = 11.26;  $p < 0.001$ ), but the aggregation of EP results is pointless in this case as the groups are very dissimilar. In the case of *so*, the patterns of use are not consistent across source languages (X-squared = 7.33;  $p = 0.03$ ), as interpreters working from Spanish appear to use *so* less frequently than those interpreting from Italian<sup>5</sup>.

As far as functions are concerned, *so* is predominantly used in all sub-corpora to draw conclusions and inferences and to present a course of action whose need is motivated by the preceding segment(s) (example (5) from the BNC). Consequential relationships are quite frequent, especially in the interpreting corpus, as illustrated in (6).

(5) my honourable friend I think wanted to intervene with a further thought on this point  
**so** I give way to her [BNC JSF]

(6) a society where technological advances enable us to organise thing in time things  
in time and space differently // **so** everybody on this planet can find his or her place  
[EPICG\_2008.08.10\_formalsitting\_betancourt\_en]

In the interpreting, *so* is also regularly used to mark the end of the turn/discourse, as in (7).

(7) most of these things in their spirit at least correspond to the decision taken by the JHA  
Council on the fifth and sixth of June / **so** madame chairman those are a few comments  
that I wanted to give the House [EPICG\_2008.01.09\_reseau\_judiciaire\_jacquesbarrot\_I\_en]

Examples of this type of closure are also found in the BNC, but they are marginal.

The reprise functions of *so*, marking a return to the main argument, a repetition, a reformulation or a list of examples, occur in all sub-corpora. Surprisingly, *so* is also used quite often to mark a shift of topic, as illustrated in (8).

5 It may be argued that it is very likely that English interpretations both from Spanish and Italian are carried out by the same interpreters or at least that there is partial overlap. In that case the formulation should be adapted.

- (8) I would like to congratulate Vladimir Špidla who has always supported this initiative and who accepted not to withdraw this text even when there was a stalemate in Brussels years ago/ and it was during the Slovenian presidency that an agreement was reached / *so* I would like to highlight the efforts made by the United Kingdom and I would like to pay tribute to the social partners agreement [EPICG \_\_2008.20\_\_10\_\_Travail Interimaire\_\_bertrandxavier\_\_I\_\_en]

Topic shifting is not found in the non-interpreted corpora, where this particular function is usually held by *now*. Finally, *so* is also used by interpreters as a hesitation marker. In 52% of its occurrences in interpretation, *so* can be presumed to be triggered by a source item.

#### 4. Discussion

The picture that emerges from the data is extremely confused: every group teams up with another group at least once: MEPs and EP interpreters in the case of *well*; MEPs with British MPs with respect to *well* and *so* and British MPs and EP interpreters in the case of *now*. Except for the data on *well*, no evidence for an EP discourse community consisting of MEPs and EP interpreters can be drawn from the analyses. The data do, however, support the view that interpreters alone act as a discourse community: discourse marker use is largely consistent across source languages. The likelihood that the same interpreter(s) is/are recorded for these source languages is small as the French>English data set is drawn from plenary sessions held in 2008, while the Spanish>English and Italian>English data sets are drawn from sessions four years earlier.

Another fairly consistent pattern that can be found in the data is that interpreters use discourse markers more frequently than MEPs. This seems to indicate that the modal divide between oral production based on oral input (interpreters) and oral production based on written input (MEPs) plays a role. As mentioned before, MEPs frequently read written statements they have prepared in advance. Discourse markers are known to be more frequent in spoken registers than in written. A written preparation is, therefore, likely to reduce the numbers of discourse markers in oral production. This seems to be confirmed by the higher frequencies of discourse markers in the BNC (S \_\_Parliament): written preparation seems to be less frequent as debates in the British Parliament are marked by direct interaction in the shape of question-answer sequences or rapid exchanges of views. Interpreters appear to use discourse markers quite autonomously. Source text triggers can only explain a third of the occurrences of discourse markers in interpretation. Autonomy is highest in the case of *well* and lowest in the case of *so*, where interpreters seem to have been prompted by source items in just over half of the cases.

With regard to functions, the findings run parallel overall, but the functional boundaries seem to be more permeable in interpreting than in non-interpreted speech: discourse markers are found in functions that are normally held by one of the other discourse markers. This could be due to the fact that fewer attentional resources are available for production in interpreting as resources are already used for the analysis of the source speech.

Surprisingly, the data also confirm tendencies observed in Cartoni *et al.* (2011). Using the Europarl parallel corpus to test lexical homogeneity of the comparable components of the corpus, Cartoni *et al.* (*ibid.*) reach the conclusion that homogeneity is surprisingly high between translated and non-translated text. This is completely in line with the aforementioned findings from lexical studies on interpreting (Sandrelli/Bendazzoli 2005; Ferraresi/Miličević 2017). However, Cartoni *et al.* (2011) also found that patterns of discourse marker use vary considerably between translated and non-translated texts. The discourse markers they investigate are a different set (causal connectives) than the ones investigated here, but the findings run surprisingly parallel: discourse markers appear to vary more across genres than crude lexical properties. Cartoni *et al.* (*ibid.*) conclude that discourse markers are, therefore, better discriminators in comparable studies.

## 5. Conclusions

The main question of this study was: what is the risk of using comparable components of parallel interpreting corpora, if we know that these comparable components are drawn from sessions in which the groups generating the data share a working environment and could therefore influence one another? This question was inspired by contradicting and inconclusive findings often found in corpus-based studies of EP interpreting. These findings suggest that non-interpreted and interpreted spoken registers in the EP differ less than expected.

To investigate the potential linguistic convergence that is likely to result from this mutual influence, both a theoretical and an empirical approach were taken. The theoretical approach sought to determine whether EP interpreters and MEPs could be analysed as constituting one single discourse community, according to the criteria put forward by Swales (1990). It turned out that this was the case.

The empirical study, based on three discourse markers, showed that there was no consistent linguistic conversion between MEPs and EP interpreters as opposed to British MPs. The existence of an EP discourse community comprising both members of parliament and interpreters could, therefore, not be confirmed for the investigated items. The contradiction between results based on an analysis of discourse marker use, on the one hand, and results based on the analysis of various lexical variables, on the other, is not unique. A similar case was reported in Cartoni *et al.* (2011) regarding comparable translated and non-translated components of a parallel corpus of EP translations.

At face value, this means that the risk attached to comparable investigations of parallel interpreted data drawn from the EP is probably not so high. Which brings us back to the fundamental question as to why analyses of lexical variables often yield inconclusive results, especially when they are compared with analyses of these variables in translation. The fact that analyses based on EP translation also fail to produce significant results (Cartoni *et al.* 2011) is intriguing and suggests that these patterns are somehow connected to an EP discourse anyway.

## References

- Aijmer K. (2013) *Understanding Pragmatic Markers: A Variational Pragmatic Approach*, Edinburgh, Edinburgh University Press.
- Aijmer K. / Foolen A. / Simon-Vandenberghe A.-M. (2006) "Discourse particles in the perspective of heteroglossia", in K. Fischer (ed.) *Approaches to Discourse Particles*, Oxford/Amsterdam, Elsevier, 101–114.
- Baker M. (1993) "Corpus Linguistics and Translation Studies: implications and applications", in M. Baker / G. Francis / E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, Amsterdam/Philadelphia, John Benjamins, 233–250.
- Bartłomiejczyk M. (2017) "The interpreter's visibility in the European Parliament", *Interpreting* 19/2, 159–185.
- Bernardini S. / Ferraresi A. / Miličević M. (2016) "From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective", *Target* 28/1, 61–86.
- Bernardini S. / Ferraresi A. / Russo M. / Collard C. / Defrancq B. (forthcoming) "Building interpreting and intermodal corpora: a how-to for a formidable task", in C. Bendazzoli / M. Russo / B. Defrancq (eds) *Making Way in Corpus-based Interpreting Studies*. Springer.
- Biber D. (2006) *University Language: A Corpus-based Study of Spoken and Written Registers*, Amsterdam/Philadelphia, John Benjamins.
- Blakemore D. (2002) *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*, Cambridge, Cambridge University Press.
- Blakemore D. / Gallai F. (2014) "Discourse markers in free indirect style and interpreting", *Journal of Pragmatics* 60, 106–120.
- Bolden G. (2006) "Little words that matter: discourse markers "so" and "oh" and the doing of other-attentiveness in social interaction", *Journal of Communication* 56/4, 661–688.
- Buysse L. (2012) "So as a multifunctional discourse marker in native and learner speech", *Journal of Pragmatics*, 44/3, 1764–1782.
- Calzada-Pérez M. (2007) *Transitivity in Translating: The Interdependence of Texture and Context*, Bern, Peter Lang.
- Cartoni B. / Zufferey S. / Meyer T. / Popescu-Belis A. (2011) "How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives", in Association for Computational Linguistics (ed.) *Proceedings of the 4th Workshop on Building and Using Comparable Corpora, 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 24 June 2011*, 78–86.
- Defrancq B. (2016) "Well, interpreters... a corpus-based study of a pragmatic particle used by simultaneous interpreters", in G. Corpas Pastor / M. Seghiri (eds) *Corpus-based Approaches to Translation and Interpreting*, Bern, Peter Lang.
- Defrancq B. / Plevoets K. / Magnifico C. (2015). "Connective markers in interpreting and translation: Where do they come from?", in J. Romero Trillo (ed.) *Yearbook of Corpus Linguistics and Pragmatics 2015*, Bern, Springer, 195–222.

- Duflou V. (2014) *Be(com)ing a Conference Interpreter. An Ethnography of EU Interpreters as a Professional Community*, PhD thesis, KULeuven.
- European Communities (2003) *Joint Practical Guide of the European Parliament, the Council and the Commission for persons involved in the drafting of legislation within the Community institutions*, <[http://www.eesc.europa.eu/resources/docs/joint\\_practical\\_guide\\_en.pdf](http://www.eesc.europa.eu/resources/docs/joint_practical_guide_en.pdf)> .
- Ferraresi A. / Miličević M. (2017) "Phraseological patterns in translation and interpreting. Similar or different?" in G. De Sutter / M.-A. Lefer / I. De-laere (eds) *Empirical Translation Studies. New Methodological and Theoretical Traditions*, Berlin, Mouton-De Gruyter, 157-182.
- Giles H. / Coupland N. / Coupland J. (1991) "Accommodation Theory: communication, context, and consequence", in H. Giles / J. Coupland / N. Coupland (eds) *Contexts of Accommodation*, New York, Cambridge University Press, 1-68.
- Hale S. (2004) *The Discourse of Court Interpreting*, Amsterdam/Philadelphia, John Benjamins.
- Halliday M. (1994) *An Introduction to Functional Grammar. Second Edition*, London, Edward Arnold.
- Johansson S. / Ebeling J. / Oksefjell S. (1999-2002) *English-Norwegian Parallel Corpus: Manual*, Oslo, Department of British and American Studies, <<https://www.hf.uio.no/ilos/english/services/omc/enpc/ENPCmanual.pdf>>.
- Kajzer-Wietrzny M. (2012) *Interpreting Universals and Interpreting Style*, PhD thesis, University of Poznan, <<https://repozytorium.amu.edu.pl/jspui/bitstream/10593/2425/1/Paca%20doktorska%20Marty%20Kajzer-Wietrzny.pdf>>.
- Kajzer-Wietrzny M. (2015) "Simplification in interpreting and translation", *Across Languages and Cultures* 16/2, 233-255.
- Lam P. (2009) "The effect of text type on the use of *so* as a discourse particle", *Discourse Studies* 11/3, 353-372.
- Lassen I. (2003) *Accessibility and Acceptability in Technical Manuals: A Survey of Style and Grammatical Metaphor*, Amsterdam/Philadelphia, John Benjamins.
- Leech G. (2007) "New resources, or just better old ones? The Holy Grail of representativeness", in M. Hundt / N. Nesselhauf / C. Biewer (eds) *Corpus Linguistics and the Web*, Amsterdam, Rodopi, 133-149.
- Mason M. (2008) *Courtroom Interpreting*, Lanham, University Press of America.
- McEnery T. / Xiao R. / Tono Y. (2006) *Corpus-based Language Studies. An Advanced Resource Book*, London/New York, Routledge.
- Olohan M. / Baker M. (2000) "Reporting that in translated English: evidence for subconscious processes of explicitation?", *Across Languages and Cultures* 1/2, 141-158.
- Pettel S. (2005) *The Development of Competence in French Interlanguage Pragmatics: The Case of the Discourse Marker 'donc'*, PhD thesis, Texas University.
- Raymond G. (2004) "Prompting action: the stand-alone *so* in ordinary conversation", *Research on Language and Social Interaction* 37/2, 185-218.
- Robinson W. (2014) "Translating legislation: the European Union experience", *The Theory and Practice of Legislation* 2/2, 185-210.

- Russo M. / Bendazzoli C. / Sandrelli A. (2006) "Looking for lexical patterns in a trilingual corpus of source and interpreted speeches. Extended analysis of EPIC (European Parliament Interpreting Corpus)", *Forum* 4/1, 221-254.
- Sandrelli A. / Bendazzoli C. (2005) "Lexical patterns in simultaneous interpreting: a preliminary investigation of EPIC (European Parliament Interpreting Corpus)". *Proceedings from the Corpus Linguistics Conference Series*, <<http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/ContrastiveCorpusLinguistics/lexicalpatternsinsimultaneousinterpreting.doc>> .
- Schiffrin C. (1987) *Discourse Markers*, Cambridge, Cambridge University Press.
- Shlesinger M. (2009) "Towards a definition of interpretese: an intermodal, corpus-based study", in G. Hansen / A. Chesterman / H. Gerzymisch-Arbogast (eds) *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*, Amsterdam/Philadelphia, John Benjamins, 237-253.
- Shlesinger M. / Ordan N. (2012) "More spoken or more translated? Exploring a known unknown of simultaneous interpreting", *Target* 24/1, 43-60.
- Swales J. (1990) *Genre Analysis: English in Academic and Research Settings*, Cambridge, Cambridge University Press.
- Wenger E. (1998) *Communities of Practice: Learning, Meaning and Identity*, Cambridge, Cambridge University Press.