

4 Cross-Linguistic Variation in Spoken Discourse Markers

Distribution, Functions, and Domains

*Liesbeth Degand, Zoé Broisson, Ludivine Crible,
and Karolina Grzech*

4.1 Introduction

According to Du Bois (1985: 363), “Grammars code best what speakers do most.” What speakers do a lot is use discourse markers (DMs). In spontaneous conversation, speakers use approximately one DM every twelve words in French and one DM every eighteen words in English (Crible 2018). Yet, despite an impressive quantity of work starting in the early 1980s, the linguistic description of DMs remains scattered, first and foremost because it is a very heterogeneous linguistic category, fulfilling many different functions in discourse (for recent overviews, see Maschler and Schiffrin 2015; Blühdom et al. 2017). This has led to a multitude of case studies on particular DMs in a diversity of languages but much less to a categorical description of DMs as a linguistic class (some exceptions are Fraser 1999; Fischer 2014; Crible 2018). To gain deeper insight into this complex linguistic category, we believe more systematic cross-linguistic work is needed based on authentic data. Indeed, corpora reflect best how we use language in different situational settings, and corpora of spontaneous speech are probably the most accurate outcome of the cognitive processes underlying language production. Thus, looking at *how* we use DMs is a crucial step in finding out *why* we use them. This is where discourse annotation comes into play as a method to enrich corpus data with linguistic and extra-linguistic information in order to uncover patterns of language use. While some of these patterns are language-specific, others are shared cross-linguistically, thus informing of inherent linguistic properties. In this chapter, we present a functional annotation scheme of DMs (Crible and Degand 2019a) on the basis of which we analyze the variation in use and functions of a broad bottom-up selection of DMs across four languages from different typological families, namely French and Spanish (Romance), English (Germanic), and Polish (Slavic). Results indicate that the multilingual

annotation scheme may be applied validly to the four different languages. This makes it possible to uncover both similarities and divergences in the functional and semantic distribution of DMs.

4.2 Annotating DMs in Spoken Languages

Discourse annotation consists in enriching corpus data with information regarding phenomena beyond the clause. Thus, depending on the research question, the need to locate and extract the relevant data before analyzing and interpreting it may arise. Examples of discourse phenomena for which annotation schemes have been developed include information structure (e.g., Calhoun et al. 2005), argument structure (Marcus et al. 1993, Palmer et al. 2005), speech acts (Kirk 2016), reference (e.g., Garside 1993; Stede 2016), fluency, and disfluency cues (e.g., Shriberg 1994; Crible et al. 2019), but the most prolific work is in the annotation of discourse relations and their markers. Many different taxonomies exist, operationalized in as many discourse annotation schemes, mainly for written (Carlson et al. 2001; Prasad et al. 2008; Afantenos et al. 2012; Scholman et al. 2016) but also for spoken data (Tonelli et al. 2010; Crible and Degand 2019a). Reliable DM annotation is a highly challenging task (Spooren and Degand 2010; Stede and Pelszus 2012; Halbe 2013; Taboada and Das 2013; Crible and Degand 2019b). This difficulty is linked both to the inherent complexity of discourse annotation and to the linguistic phenomenon under scrutiny, here the discourse marker category itself.

Ideally, discourse annotation schemes should meet four different objectives in order to be of benefit for linguistic analysis. They should be (1) fine-grained; (2) reliable; (3) theoretically sound; and (4) cognitively realistic. In other words, the ideal discourse annotation scheme provides information that is precise and fine-grained enough to allow for detailed linguistic analyses. It is also designed in view of maximal reliability warranting replicable and unbiased results and allowing objective comparison when the model is applied to different types of data (different languages, different registers), thus promoting reusability. Basing the annotation scheme on theoretically sound foundations has the advantage of bringing empirical evidence to theoretical findings, thus strengthening our linguistic models and theories. Finally, in order to bridge the gap between corpus-based work and psycholinguistic research on language processing, a cognitively realistic annotation scheme will favor fruitful multi-disciplinary and multi-method studies. The aim is thus to find the balance between methodological elegance and cognitive soundness (see Schmid 2012: 4–5).

Discourse annotation, and specifically DM annotation, also faces the specific challenges of working with spoken language. While written DM annotation,

usually referred to as (discourse) connective annotation (e.g., Prasad et al. 2008), is generally restricted to the annotation of lexical cues marking discourse relations, spoken DM annotation involves a wider class, both in terms of (syntactic) forms (conjunctions, adverbs, prepositional phrases, verbal phrases, interjections, and more) and their spectrum of functions, such as topic-related functions, interpersonal functions, or speaker-oriented functions, reflecting the speaker's difficulty in the production process. In general, the functional spectrum of DMs is larger in spoken language and there is more often simultaneous multifunctionality (Crible and Cuenca 2017). Speech is also characterized by contextual and co-textual features complicating the annotator's task. Among these, there are interruptions, missing context, inaudible context, bleached meaning, and more ambiguity in general. Yet, since spoken language is not inherently different from written language (or vice versa), we believe it is important that any annotation model of spoken language be comparable with models for written data.

4.3 A Two-Dimensional Account of Discourse Markers

The model we use in this study was originally developed for spoken French and then extended to English. An up-to-date presentation is given in Crible and Degand (2019a, 2019b).¹ Here, we focus on the model's main characteristics.

First, the annotation model follows a marker-based approach. In other words, the focus of the annotation process is on the linguistic expressions that function as DMs, not the relations underlying them. Such a functional description defines DMs as a grammatically heterogeneous, syntactically optional, poly-functional type of pragmatic marker. Their specificity is to function on a meta-discursive level as procedural cues to constrain the interpretation of the host unit in a co-built representation of ongoing discourse. They do so by either signaling a discourse relation between the host unit and its context, making the structural sequencing of discourse segments explicit, expressing the speaker's meta-comment on their phrasing, or contributing to the speaker–hearer relationship (Crible 2018: 62).

As such, the annotation model is meant to cover the whole category of DMs, accounting for its full functional spectrum. This makes it suitable for categorical studies, which investigate the behavior of DMs as a linguistic category, but less so for case studies, which tease apart all the subtle meaning differences of a specific DM. Identification of DMs in the data is determined by a number of core features, namely syntactic optionality, procedural meaning, invariability, and discourse-level scope (see, e.g., Crible 2018).

¹ The present study used a slightly different version of the scheme as it was in its development phase.

To account for the functional spectrum of DMs, the model proposes two independent dimensions of analysis. On the one hand, DMs are coded according to the discourse domain in which they work, corresponding to the speaker's communicative intention. The question at stake is: What aspect of discourse is the marker targeting? Is it relating events in the world (ideational domain), linking mental states or attitudes (rhetorical domain), structuring the discourse (sequential domain), or managing speaker–hearer interaction (interpersonal domain)? Independently from the domain annotation, DMs are attributed a discourse function corresponding to the specific contribution of the marker to the given context (one of fifteen functions among which [consequence], [addition], [temporal], [specification], [contrast], ...). Examples (1) through (4) illustrate how the DM *donc* 'so' can be used to express the function [consequence] in the four domains.

- (1) *euh Dreyfus donc euh a valu on va dire à Zola euh de s/ d'émigrer en en Angleterre pourquoi parce qu'on on lui a reproché son intervention peut-être un peu trop radicale euh et donc Zola euh Zola va devoir partir euh parce que peut-être a-t-il été trop franc*
uh Dreyfus caused let's say Zola uh to emigrate in England why because he was criticized for his intervention maybe a little too radical uh and **donc** 'so' Zola uh Zola will have to leave uh because maybe he was too frank [ideational domain] (LOCAS-F, conv-f_2)
- (2) *c'est pas évident quand euh on n'a que le _ que le CESS _ donc euh je me dis que peut-être euh à la recherche d'une euh d'une formation*
'it's not easy when uh you only you only went to high school **donc** 'so' uh I think that maybe uh looking for some training [rhetorical domain] (LOCAS-F, conv-f_1)
- (3) *et donc voilà donc euh _ suite à ça ben j'avais con/ j'ai continué les cours et puis euh*
and so that's it **donc** 'so' uh after this well I took the classes again and then uh [sequential domain] (LOCAS-F, conv-f_1)
- (4) *<speaker1> euh enfin je n- ça ne me convenait pas donc euh <speaker2> et qu'est-ce qui s'est passé?*
<speaker1> uh well I it didn't work for me **donc** 'so' uh <speaker2> and what happened? [interpersonal domain] (LOCAS-F, conv-f_1)

In (1), the fact that Zola moved to England is the direct, factual consequence of his involvement in Dreyfus's case. In (2) the speaker concludes that she should look for training on the basis of a fact (she does not have a superior degree), and this conclusion uses evaluative language ("maybe," "I think") and hesitation markers ("uh"), which testifies of its epistemic, rhetorical nature. In (3), *donc* is used in the context of hesitations and helps the speaker restart after a short interruption, taking up her previous narrative. In (4), the consequence is left

open, to be reconstructed by the other speaker, as signaled by the turn-final position and the suspensive intonation.

This model combines the strengths of attested semantically driven taxonomies (see Prasad et al. 2014) with the flexibility of domain or discourse-level oriented approaches (Schiffrin 1987; Sweetser 1990). It allows a simultaneous account in terms of polyfunctionality (i.e., the fact that a marker can be used in different domains) and polysemy (i.e., the fact that a marker can express different senses), two aspects that are intrinsically bound to the use of DMs in context. This aligns with Schiffrin's (1987: 325) seminal observation that

[i]t is because discourse is multiply structured, and its various components integrated with each other, that multiple relations hold between utterances – not because markers themselves realize a different function (one devoted to ideas, one to action, and so on) with each occasion of use. (Schiffrin 1987: 325)

Theoretically, every domain-function combination is possible, yet systematic corpus analysis should make out which combinations are attested across languages and registers.

4.4 French, Spanish, English, and Polish Discourse in Contrast

The taxonomy with two independent levels of analysis was applied to spoken unplanned dialogues in four languages from three typologically distinct families: French and Spanish (Romance), English (Germanic), and Polish (Slavic). For each language, we annotated a comparable dataset of approximately 30 minutes of conversational speech, amounting to 5,000–6,000 words. Data for English was extracted from the British component of the International Corpus of English (ICE-GB) (Nelson et al. 2002). For French, we used the Louvain Corpus of Annotated Speech–French (LoCAS-F) (Degand et al. 2014), with recordings from France and Belgium. Spanish (Peninsular) data came from the spoken component of the Backbone corpus (~~Backbone~~) (Kohn 2012) and the Polish data from the spoken corpus from the Polish CLARIN infrastructure (~~CLARIN~~) (Pęzik 2015). Annotators used sound-aligned transcripts and the sound was available during annotation. Both French and Polish were double-blind annotated, while the English and Spanish data were analyzed by only one annotator.

The first step in the coding process consists in identifying the DMs in the four datasets following the criteria mentioned in Section 4.3: syntactic optionality, procedural meaning, invariability, and discourse-level scope. This step resulted in the identification of 286 DMs in English (30 types), 442 DMs in French (35 types), 690 DMs in Spanish (71 types), and 847 DMs in Polish (79 types). All segments containing a DM were extracted with context and were assigned one domain and one function, making use of an operationalized coding scheme.

Table 4.1 *Taxonomy of domains and functions (Crible and Degand 2019b)*

ideational	rhetorical	sequential	interpersonal
[addition] [alternative] [cause] [concession] [condition] [consequence] [contrast] [punctuation] [specification] [temporal] [topic]			

For this study, the taxonomy used is presented in Table 4.1, based on Crible and Degand (2019b). In view of the results of the present contrastive analysis, it was slightly revised in later work (Crible and Degand 2019a).

A total of 442 DMs were consensually identified by 2 annotators in the French dataset. After the first annotation training phase, agreement on domains was 71 percent (314/442) and 78 percent (343/443) for functions. Thus, it appears that functions are easier to code, since they are closer to the DMs' core meaning. After discussion of all disagreements, a revised operational coding scheme was established that led to an improved agreement of 84 percent for the domains and 83 percent for the functions. This coding scheme was used for the four languages under scrutiny.

4.4.1 *French Discourse Markers in Spontaneous Interaction*

Table 4.2 illustrates that DM use is unevenly distributed over domains and functions in the French dataset, with 50 percent of the DMs being used in the sequential domain, followed by the rhetorical domain (30.5 percent) and the ideational (9.9 percent) and interpersonal (9.5 percent) domains at some fair distance. The imbalance is also present in the frequency of functions fulfilled by the DMs. All functions appear to be distributed unevenly over the four domains, and most of them have a privileged domain in which they come to expression (highlighted in Table 4.2). Strikingly, the function [condition] was not used in this subcorpus of spontaneous face-to-face interaction, while the functions [contrast] and [topic] are hardly used and occur only in the ideational and sequential domains, respectively.

A further imbalance in the data comes from the highly diverging frequencies with which the different functions are expressed by means of DMs. In French spontaneous interaction, the most frequent DM function is [punctuation] amounting to 26 percent of the DM uses. In other words, in a quarter of the cases speakers use DMs in a semantically vaguely defined way as “discourse punctuators,” corresponding to a kind of typographical comma, as illustrated with the use of *hein* ‘right’ in example (5).

Table 4.2 *Distribution of DMs across domains and functions in the French dataset (dominant functions are highlighted)*

Function	Domain				Total
	Ideational	Rhetorical	Sequential	Interpersonal	
punctuation	0	3	78	33	114
addition	8	10	65	0	83
concession	1	51	22	7	81
consequence	6	31	12	1	50
alternative	3	11	21	0	35
specification	0	13	12	0	25
temporal	22	0	2	0	24
cause	0	16	0	1	17
topic	0	0	9	0	9
contrast	4	0	0	0	4
condition	0	0	0	0	0
Total	44	135	221	42	442

- (5) *j'avais déjà hésité **hein** donc entre euh enfin entre institutrice primaire ou euh GRH (LOCAS-F, conv-f_1)*
 I had already hesitated hein ('right') donc ('well') over uhm enfin ('like') over primary school teacher and uhm HR

Further qualitative analysis of the data shows that the expression of the function [punctuation] actually comes on the account of four (nearly) monosemous DMs, adding up together to 84 percent of the DM spectrum for this function, namely the already mentioned *hein* (17/17), *quoi* (17/17), *bon* (36/36), and *ben* (26/28) (see Table 4.3). In other words, French seems to have developed specific semantically vague DMs that are used to punctuate speakers' discourse, thus indicating their will to hold the floor, weakly structuring their discourse while planning. Comparison with the other languages will make clear whether this is a specific idiosyncratic French phenomenon or whether it might be specific to the register under analysis.

Table 4.3 displays the DM types used to express the different DM functions in the taxonomy. The most frequent DM types per function are highlighted in bold, and DMs expressing more than one function are underlined. It clearly appears that there is no one-to-one mapping between DM types and the functions they express, even if most markers seem to be more dedicated to expressing one function depending on the core meaning they express. Thus, *et* 'and' is a typical additional marker, *mais* 'but' a typical concessive marker, and *donc* 'so' is most often used to express consequence. This by no means makes these markers monosemous, however. Notwithstanding the fairly limited

Table 4.3 *Distribution of DM types over discourse functions in the French dataset*

Function	DM types
punctuation	<i>bon, ben, quoi, hein</i> , voilà, beh, eh ben, tu vois, <u>alors</u> , eh bien, là, allez
addition	<u>et, puis, alors, en plus, même</u>
concession	<i>mais</i> , <u>en fait, pourtant, bien que, quoique, même, en même temps, quand même</u> , encore que, après, maintenant
consequence	<u>donec, alors, enfin</u>
alternative	<u>enfin, ou, alors</u>
temporal	<u>puis, et, alors</u>
cause	<u>parce que</u>
specification	<u>enfin, mais, en fait, donc, ben, alors</u>
topic	<u>et, alors, donc, puis</u>
contrast	<u>mais</u>

dataset, it is noteworthy to observe that only three functions ([punctuation], [concession], [cause]) attract mainly specific (monosemous) DM types, while the other functions are expressed mainly by means of polysemous DM types (underlined in Table 4.3).

Taking into account DM types with a raw frequency of ten and more (covering 90.7 percent of the tokens), Figure 4.1 suggests that high frequency goes hand in hand with high polyfunctionality (use in more than one discourse domain), with the exception of the interpersonal DM *hein* ‘right.’ Whether this correlation holds cross-linguistically needs to be confirmed in the remainder of this study.

4.4.2 Spanish Discourse Markers in Spontaneous Interaction

The second Romance language under investigation in this study is Spanish. As can be observed in Table 4.4, DM tokens are also unevenly distributed across both domains and functions in this sample. Similarly to French, the sequential domain is the most represented, with 39.1 percent of the tokens. Conversely, however, the interpersonal domain is more frequent and ranks second, with 30.6 percent of all tokens, followed by rhetorical (19.8 percent) and ideational DMs (10.4 percent). The interpersonal domain thus appears to play a more prominent role in Spanish, which suggests that DMs in conversational Spanish may be more diversified than in French and are often used as a means to manage the interactivity of an exchange besides structuring it.

Zooming in, functions also appear to be unevenly distributed, and this both generally in the sample and across the four domains. Seven functions show a preference for one domain over the others (highlighted in Table 4.4). Once again, [punctuation] is found to be the most frequent function, representing

Table 4.4 *Distribution of DMs across domains and functions in the Spanish dataset*

Function	Domain				Total
	Ideational	Rhetorical	Sequential	Interpersonal	
Punctuation	0	0	87	148	235
Addition	12	14	66	19	111
Specification	7	48	23	21	99
Consequence	10	23	9	8	50
Topic	0	1	47	0	48
Concession	6	24	5	4	39
Temporal	6	10	9	4	29
Contrast	14	5	6	3	28
Cause	6	10	4	4	24
Alternative	8	0	10	0	18
Condition	3	2	4	0	9
Total	72	137	270	211	690

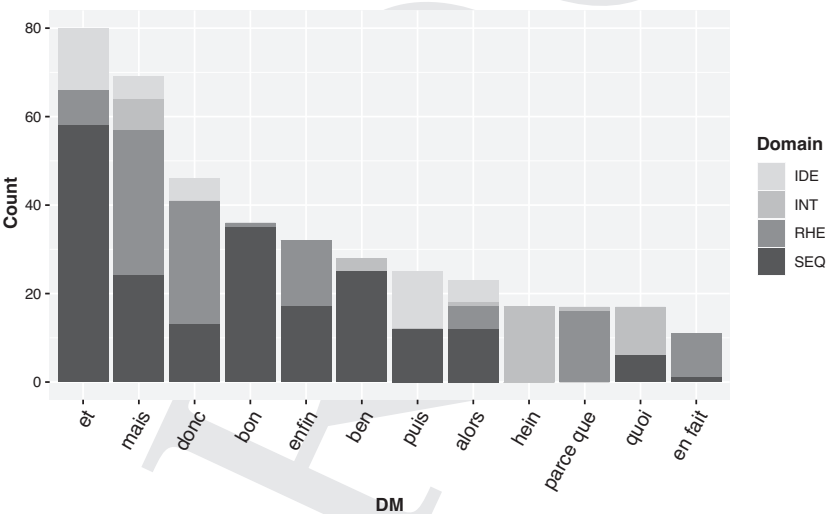


Figure 4.1 Frequency and polyfunctionality (domains) of DM use in the French dataset

approximately a third of the total number of identified DMs (34 percent). More than half of its instances are associated with the interpersonal domain, as illustrated in example (6).

92 Liesbeth Degand, Zoé Broisson, Ludivine Crible, and Karolina Grzech

- (6) No, **hombre**, después de la carrera supongo que tendré que hacer la oposición (Backbone bb_es021)
No, **hombre** ('man'), after my degree I suppose I'll have to do teacher training

The most frequent functions of Spanish DMs are the same as in the French data (i.e., [addition], [concession], and [consequence]). In the Spanish sample, [specification] DMs are better represented and seem to be particularly typical of the rhetorical domain, such as in example (7).

- (7) ... la oportunidad de viajar, de conocer otros investigadores, otros sitios, **es decir**, otras experiencias, ¿no? (Backbone bb_es012)
... the opportunity to travel, to meet other researchers, to discover new places, **in other words** new experiences, right?

When examining frequent discourse functions in the Spanish data, [punctuation] and [specification] stand out as functions that come in many different DM types, as reported in Table 4.5. While *pues*, *no*, and *bueno* are not restricted to [punctuation], *por ejemplo* and *es decir* are typical as well as exclusive representatives of [specification]. Here, the case of [punctuation] is particularly interesting, as DMs labeled with this function tend to be monosemous in French (and in English, as will be described in Section 4.4.3). Functions that express basic discourse relations such as [addition] and [consequence] are mainly

Table 4.5 Distribution of DM types over discourse functions in the Spanish dataset

Function	DM types
Punctuation	<i>pues</i> , <i>no</i> , <i>bueno</i> , <i>la verdad</i> , <i>muy bien</i> , <i>claro</i> , <i>y</i> , <i>en fin</i> , <i>ya</i> , <i>y demás</i> , <i>bien</i> , <i>hombre</i> , <i>no sé</i> , <i>nada</i> , <i>o</i> , <i>ahora</i> , <i>a lo mejor</i> , <i>a ver</i> , <i>de acuerdo</i> , <i>entonces</i> , <i>perfecto</i> , <i>pero</i> , <i>por cierto</i> , <i>por lo menos</i> , <i>por supuesto</i> , <i>sí</i> , <i>sobre todo</i> , <i>vale</i> , <i>vamos</i>
Addition	<i>y</i> , <i>también</i> , <i>además</i> , <i>de hecho</i> , <i>pues</i> , <i>bueno</i> , <i>en verdad</i> , <i>la verdad</i> , <i>luego</i> , <i>pero después</i> , <i>sobre todo</i> , <i>tampoco</i>
Specification	<i>por ejemplo</i> , <i>es decir</i> , <i>y</i> , <i>la verdad</i> , <i>a lo mejor</i> , <i>o sea</i> , <i>pues</i> , <i>a ver</i> , <i>así</i> , <i>no</i> , <i>o</i> , <i>sí</i> , <i>a mí</i> , <i>a nivel</i> , <i>bueno</i> , <i>en fin</i> , <i>más bien</i> , <i>no sé</i> , <i>quizá</i> , <i>sobre todo</i>
Consequence	<i>entonces</i> , <i>y</i> , <i>pues</i> , <i>así</i> , <i>para luego</i> , <i>a mí</i> , <i>de hecho</i> , <i>por consiguiente</i> , <i>por eso</i> , <i>por lo tanto</i>
Topic	<i>y</i> , <i>pero</i> , <i>ahora</i> , <i>por otro lado</i> , <i>pero luego</i>
Concession	<i>pero</i> , <i>aunque</i> , <i>aún</i> , <i>ahí</i> , <i>entonces</i> , <i>no</i> , <i>sino</i> , <i>y</i>
Temporal	<i>y</i> , <i>ahora (mismo)</i> , <i>a la hora</i> , <i>hoy en día</i> , <i>luego</i> , <i>siempre</i> , <i>cada vez</i> , <i>de momento</i> , <i>en principio</i> , <i>entonces</i> , <i>hoy</i> , <i>mañana</i> , <i>pronto</i> , <i>una vez que</i>
Contrast	<i>pero</i> , <i>y</i> , <i>sin embargo</i> , <i>sino</i>
Cause	<i>porque</i> , <i>ya que</i> , <i>por lo que</i> , <i>luego</i>
Alternative	<i>o</i> , <i>o bien</i>
Conditional	<i>si</i>

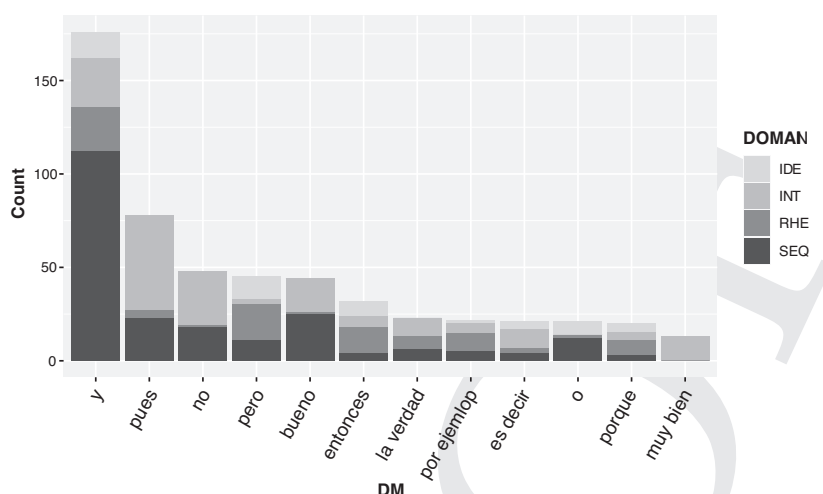


Figure 4.2 Frequency and polyfunctionality of DM use in the Spanish dataset

expressed by one DM type. Generally, the trends uncovered in the use of Spanish DMs coincide with those found for French and further suggest that there is no one-to-one mapping between DM types and functions.

We next focus on the twelve DM types with a raw frequency of ten or more ($N=543$, 78.7 percent). Figure 4.2 displays signs of a positive relationship between frequency and polyfunctionality (across domains), similar to the French data. In terms of polysemy, however, most Spanish DMs express one main function only, with the notable exception of the DM *y* ‘and’, which occurs in all four domains and across a total of eight functions. At the other end of the spectrum, *muy bien* is only found as the combination of the interpersonal domain and the [punctuation] function.

Overall, this first analysis of Spanish data using Crible and Degand’s (2019b) taxonomy reveals that much like in French, Spanish markers produced in spontaneous settings operate in both relational and interactive aspects of discourse. The latter include the more speech-specific domains and functions of the taxonomy, such as the sequential and interpersonal domains and the [punctuation] function all widely represented in the data, while the former correspond to the rhetorical domain and functions such as [addition], [consequence], and [concession]. While these results highlight similarities with French when it comes to the distribution of domains and functions, several differences between the two languages are also brought to light, particularly regarding the overall degree of polysemy of their respective frequent marker types.

Table 4.6 *Distribution of DMs across domains and functions in the English dataset*

Function	Domain				Total
	Ideational	Rhetorical	Sequential	Interpersonal	
Punctuation	0	0	39	48	87
Addition	20	2	48	0	70
Concession	0	32	10	0	42
Specification	0	18	15	0	33
Consequence	1	11	3	0	15
Contrast	7	1	3	0	11
Cause	0	11	0	0	11
Alternative	3	4	0	0	7
Condition	0	4	0	0	4
Topic	0	0	4	0	4
Temporal	0	0	2	0	2
Total	31	83	124	48	286

4.4.3 *English Discourse Markers in Spontaneous Interaction*

The English data shows the same imbalance as observed in the two Romance languages and a similar preference for the sequential domain, which takes up 43.36 percent of all DMs, as can be seen in Table 4.6.

The ranking of domains is closer to Spanish than to French, with a higher proportion of interpersonal uses of DMs, mainly expressed by the highly frequent *you know*. Almost all functions can be expressed in the rhetorical and/or sequential domains, whereas the interpersonal domain is restricted to the [punctuation] function, and the ideational domain mainly consists of [addition] and [contrast] uses. As in French and Spanish, some DM functions show one prevalent domain, but the distribution of domain–function combinations seems more balanced than in Romance languages, with smaller differences between domains for [addition], [punctuation], and [specification]. Similar to the French data, the most frequent function in this English sample is [punctuation], as in example (8).

- (8) *you can't actually rely on what he's going to have always (0.473) you know you think oh I'll uh you know I've got people coming I'll get some salmon (ICE-GB S1A-010)*

Again, most DMs expressing the function of [punctuation] are monosemous, with the exception of *well*, which also shows some uses of [specification] (five cases) and [alternative] (one case). Like French, but unlike Spanish,

Table 4.7 *Distribution of DM types over discourse functions in the English dataset*

Function	DM types
punctuation	<u>well</u> , <u>you know</u> , <i>you see</i> , <i>look</i> , <i>yeah</i>
addition	<u>and</u> , <i>plus</i>
concession	<u>but</u> , <i>although</i> , <u>actually</u> , <i>though</i> , <i>even if</i>
specification	<u>I mean</u> , <u>well</u> , <u>actually</u> , <i>sort of</i> , <u>so</u> , <i>like</i> , <i>for instance</i> , <i>as it were</i> , <i>kind of</i> , <i>in fact</i>
consequence	<u>so</u> , <i>then</i>
contrast	<u>but</u> , <u>and</u>
cause	<u>because</u> , <i>insofar as</i>
alternative	<i>or</i> , <u>well</u> , <u>I mean</u>
topic	<u>and</u> , <i>now</i> , <i>anyway</i>
condition	<i>if</i>
temporal	<i>secondly</i> , <i>first of all</i>

[punctuation] tends to be expressed by semantically specialized DMs in English. The other frequent functions are quite similar to the other languages in their nature and ranking, as can be seen in Table 4.7.

Seven DMs, with ten occurrences or more, account for 58.5 percent of the data and their domain distribution is represented in Figure 4.3. In English, too, high frequency tends to be associated with high polyfunctionality (several domains) and high polysemy (several functions), with the exceptions of *you know* (exclusively interpersonal punctuation uses) and *because* (exclusively rhetorical causal uses). All other DMs in this subset express at least two domains and two functions, although there is always one dominant value. The most polyfunctional DM is *and* with three other values besides [addition]. All in all, while there is no one-to-one mapping between DMs and their functions, and DM polyfunctionality has been confirmed in this dataset, we can observe that the most frequent DM types in English are mainly used to express basic discourse relations as well as more speech-specific functions related to the management of the discourse flow and interaction.

4.4.4 Polish Discourse Markers in Spontaneous Interaction

In the Polish dataset, like in French, Spanish, and English, DMs are most often used sequentially (36.2 percent), as shown in Table 4.8. Second in line is the interpersonal domain (25.5 percent), while the ideational and rhetorical domains both take up around 19 percent of the DM use. As compared to the other languages in our dataset, the interpersonal domain is strongly represented. We do not believe that this higher frequency follows from an artifact of the data, as we were careful to work with comparable samples in the four

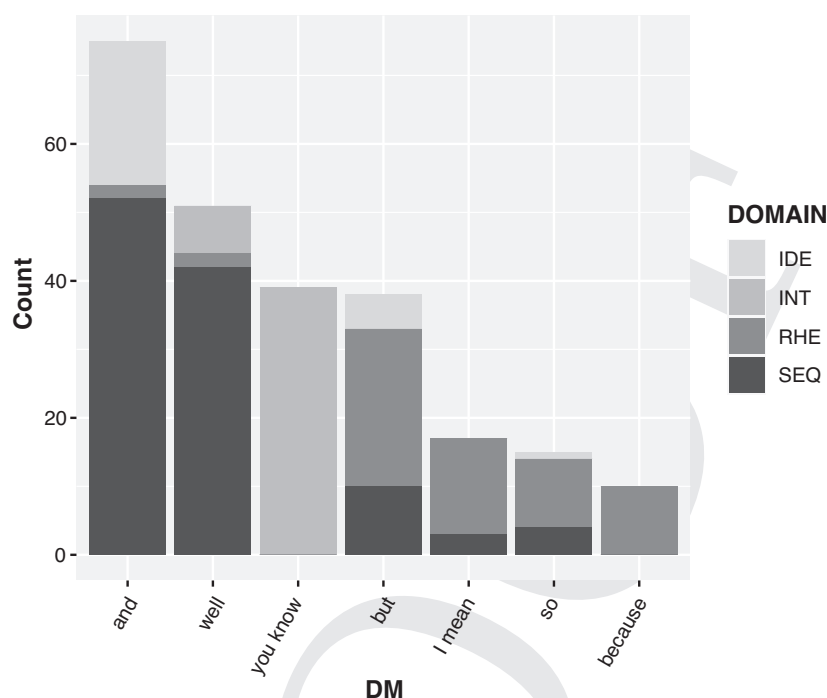


Figure 4.3 Frequency and polyfunctionality of DM use in the English dataset

languages, that is, informal, private, face-to-face conversations between two or more acquainted speech partners. Moreover, we do not think that corpus size is the main reason for this discrepancy either, since the analysis of larger datasets in English and French (Crible 2018) shows a similar weaker presence of the interpersonal domain. Possible explanations will be developed in Section 4.5.

Similar to English, only a few functions show one prevalent domain, while several others are frequent in two domains. The function of [punctuation] was the most frequent, with a large number of tokens in both the sequential and the interpersonal domains. Patterns of domain-function combinations are generally similar to the other languages in the study. All four domains of the taxonomy see the expression of a variety of functions in the Polish data. None of the domains can be said to be homogeneous with respect to the discourse functions of the DMs they cover.

Table 4.9 shows the distribution of DM types across the different functions, ordered according to their decreasing frequency in the sample. Each function is expressed through a heterogeneous set of DMs. The rich inventory of DM types

Table 4.8 *Distribution of DMs across domains and functions in the Polish dataset*

Function	Domain				Total
	Ideational	Rhetorical	Sequential	Interpersonal	
punctuation	3	3	153	144	303
addition	35	15	70	19	139
specification	18	37	35	12	102
topic	2	3	31	33	69
temporal	55	7	1	0	63
adversative*	13	32	4	8	57
cause	10	38	7	0	55
alternative	10	11	2	0	23
condition	11	10	1	0	22
consequence	3	8	3	0	14
Total	160	164	307	216	847

* [Concession] and [contrast] were merged into the [adversative] function after consideration of systematic disagreements between annotators.

within each function sets Polish apart from English and French and is more in line with the Spanish data. This may be due to the larger number of DM types attested in the analyzed sample (English: 30, French: 35, Spanish: 71, Polish: 79) and the larger number of tokens (English: 286, French: 442, Spanish: 690, Polish: 847), but this explanation is not convincing in comparison to the Spanish data.

Figure 4.4 shows how the Polish DM types, listed in Table 4.9, are distributed across the four discourse domains.

The marker type that stands out as the most frequent in the data is *wiesz* ‘you know’ (cf. Diskin-Holdaway, Chapter 9, this volume). It occurs 169 times, which constitutes almost 20 percent of all the DM tokens in the dataset. The occurrences of *wiesz* are distributed between the sequential ($N=81$) and the interpersonal ($N=88$) domain. Interestingly, the composite *wiesz co* ‘you know what’ patterns differently from *wiesz*. Rather than being polyfunctional, this marker is only attested in the interpersonal domain. On the other hand, while for *wiesz* the dominant function is [punctuation], attributed to 156 out of 169 tokens of the DM in the sample, *wiesz co* is more polysemous (several functions).

Polysemy is by no means unique to *wiesz co*. In the analyzed dataset, only two of the seventeen frequent DM types are monosemous, namely *potem* ‘then’/‘after’, *no nie* (roughly translatable as ‘you know’, but literally meaning ‘well no’). Apart from being a strictly [temporal] DM, *potem* is also only

Table 4.9 Distribution of DM types over discourse functions in the Polish dataset

Function	DM Types
punctuation	<u>wiesz</u> , <u>no</u> , <u>no nie</u> , <u>nie</u> , <u>sluchaj</u> , <u>tak</u> , <u>to</u> , <u>no to</u> , <u>wiesz co</u> , <u>a</u> , <u>po prostu</u> , <u>prawda</u> , <u>w ogóle</u> , <u>znaczy</u>
addition	<u>i</u> , <u>no</u> , <u>a</u> , <u>sluchaj</u> , <u>jeszcze</u> , <u>no to</u> , <u>bo</u> , <u>poza tym</u> , <u>aż</u> , <u>to</u> , <u>tylko</u> , <u>wiesz</u> , <u>wiesz co</u> , <u>z resztą</u> , <u>ale</u> , <u>e</u> , <u>gdzie</u> , <u>i w ogóle</u> , <u>nie dosyć że ... to jeszcze</u> , <u>no sluchaj</u> , <u>po prostu</u> , <u>tak</u> , <u>w każdym razie</u> , <u>w ogóle</u> , <u>że</u>
specification	<u>no</u> , <u>to</u> , <u>w ogóle</u> , <u>na przykład</u> , <u>to znaczy</u> , <u>wiesz</u> , <u>po prostu</u> , <u>znaczy</u> , <u>ale</u> , <u>bo</u> , <u>i</u> , <u>jeszcze</u> , <u>no to</u> , <u>wiesz co</u> , <u>a</u> , <u>i tak dalej</u> , <u>tylko</u> , <u>ale to</u> , <u>byłe by</u> , <u>czekaj</u> , <u>jak</u> , <u>nawet</u> , <u>oczywiście</u> , <u>patrz</u> , <u>poza tym</u> , <u>przecież</u> , <u>tak</u> , <u>tak że</u> , <u>tak że</u> <u>wiesz</u> , <u>w końcu</u> , <u>więc</u> , <u>właściwie</u> , <u>z resztą</u>
topic	<u>sluchaj</u> , <u>a</u> , <u>no</u> , <u>wiesz co</u> , <u>wiesz</u> , <u>to</u> , <u>ale</u> , <u>bo</u> , <u>i</u> <u>wiesz</u> , <u>na przykład</u> , <u>naprawdę</u> , <u>no i</u> , <u>no i tak</u> , <u>no sluchaj</u> , <u>no to</u> , <u>no to sluchaj</u> , <u>no</u> <u>wiesz</u> , <u>o</u> , <u>patrz</u> , <u>tak</u> , <u>tak że</u> , <u>w ogóle</u>
temporal	<u>jak</u> , <u>potem</u> , <u>jak ... to ...</u> , <u>wtedy</u> , <u>najpierw</u> , <u>jak ... no to ...</u> , <u>kiedy</u> , <u>teraz</u> , <u>a</u> , <u>co</u> , <u>co ... to ...</u> , <u>jak tylko</u> , <u>po</u> , <u>po czym</u> , <u>to</u>
adversative	<u>ale</u> , <u>a</u> , <u>przecież</u> , <u>tylko</u> , <u>a tam</u> , <u>bo</u> , <u>e</u> , <u>gdzie tam</u> , <u>nie</u> , <u>nie no</u> , <u>no</u> , <u>sluchaj</u> , <u>tylko że</u> , <u>wprawdzie</u> , <u>z tym że</u> , <u>żeby</u>
cause	<u>bo</u> , <u>dlatego że</u> , <u>to</u> , <u>żeby</u> , <u>jak ... to ...</u> , <u>ponieważ</u> , <u>więc</u>
alternative	<u>a</u> , <u>ale</u> , <u>znaczy</u> , <u>bo</u> , <u>albo</u> , <u>chyba że</u> , <u>czy</u> , <u>natomiast</u> , <u>teraz</u>
condition	<u>jak ... to ...</u> , <u>jak</u> , <u>jakby</u> , <u>to</u> , <u>a</u> , <u>chyba że</u> , <u>jak</u> , <u>jeżeli</u> , <u>tak</u>
consequence	<u>to</u> , <u>żeby</u> , <u>no to</u> , <u>tak że</u> , <u>jak ... to ...</u> , <u>no</u>

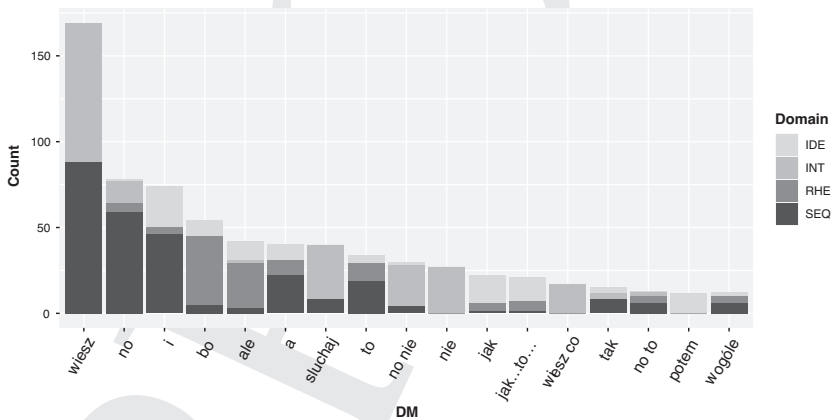


Figure 4.4 Frequency and polyfunctionality of DM use in the Polish dataset

attested in the ideational domain. Thus, it is the only DM type in the data that is not only monofunctional but also monosemous. The only other marker in the dataset that is nearly monofunctional and monosemous is *nie* ‘no’

(homophonous with the Polish negative particle). All its twenty-seven tokens occur in the interpersonal domain. Out of these, twenty-six serve the function of [punctuation] and only one serves the [adversative] function.

Comparing the polyfunctionality of the Polish DMs with that of the other languages, the strikingly high number of (frequent) DM types used interpersonally stands out. Polish counts no less than nine DM types in the interpersonal domain, while French counts six, English only two. These are all outnumbered by Spanish with eleven DM types, which are, however, used less frequently token-wise.

4.5 Measuring the Crosslinguistic Distribution of DMs across Domains and Functions

It is now time to investigate to what extent the two-dimensional annotation model makes it possible to uncover (sometimes hidden) similarities and divergences between the four languages under study. Figure 4.5 visualizes the DM domain distribution across French, Spanish, English, and Polish. In general terms, speakers across the four languages use DMs in similar ways. The most frequent domain of use is the sequential one. Thus, in spontaneous spoken face-to-face interaction, speakers use DMs in the first place to structure and organize their discourse. In the second place, DMs serve to express the speakers' rhetorical attitude and stance toward the message they are conveying. Third, speakers use DMs to maintain and manage the speaker–hearer relation during

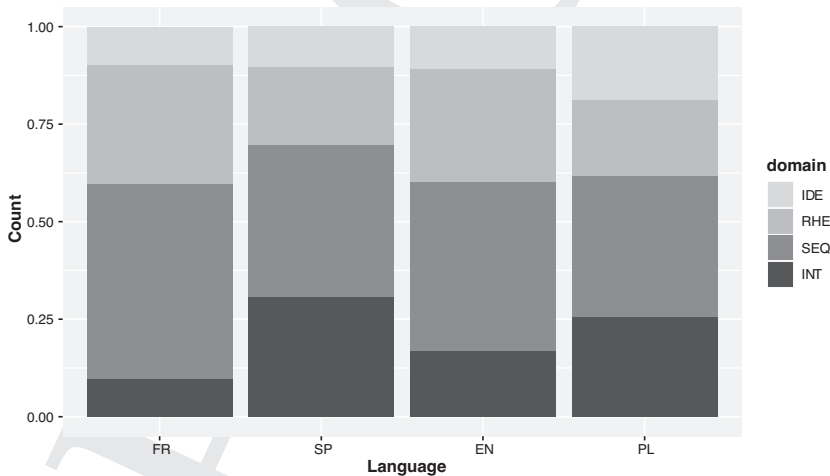


Figure 4.5 Cross-linguistic domain distribution of discourse markers

the interaction. These three general uses of DMs are all metalinguistic in nature and form the bulk of speakers' choices to make use of a DM. The least frequent use across the four languages is the ideational domain, that is, to relate discourse segments semantically and in terms of content.

Zooming in, this distribution appears to be cross-linguistically diverging in a statistically significant way ($\chi^2=126.1$, $df=9$, $p<0.001$). More precisely, an examination of the standardized residuals shows that, compared to the other languages, French stands out with its low rate of interpersonal DMs and, conversely, high frequency of rhetorical and sequential DMs. By contrast, Spanish shows a significantly higher frequency of interpersonal DMs. Polish DMs show a particularly large proportion of ideational uses and a lower proportion of rhetorical uses. The English data does not strongly differ from the other languages. DM domains reflect the speaker's communicative intentions. Given that all samples in the study correspond to very similar communicative contexts, such discrepancies in the results were not expected. Several explanations come to mind.

The first explanation is methodological. The annotation of the four languages was performed independently from one another. In other words, even if the annotation procedure was the same in the four languages, it is possible that they were not applied in exactly the same way throughout the four datasets. What is at stake here is the reliability of subtle contextual interpretations, which have been shown to be fragile when it comes to distinguishing between the sequential and the rhetorical domains as well as between the ideational and the rhetorical domains (Spooren and Degand 2010; Crible and Degand 2019b). These are precisely the domains where French and Polish show a discrepant distribution pattern. The second explanation is also methodological in nature and concerns the comparability of the datasets. Domains of DM use are tightly bound to the communicative situation (Crible 2018). Although we have tried to use comparable datasets in the four languages, it is still possible that these situations slightly diverge (e.g., more informal conversations in Polish than in French), which might have (slightly) influenced the domain distribution.

The third explanation is linguistic in nature and has to do with the linguistic potential available in the different languages to express the different domains by means of DMs. Strikingly, Spanish and Polish show a higher proportion of DMs in the interpersonal domain and, at the same time, these two languages also have a higher stock of DM types that are (frequently) used in this domain, viz. eleven and nine, respectively. We therefore suggest that the availability of DMs that are (more) typical of a specific domain influence how likely a given speaker is to use that domain. Put more simply, if a given language has more interpersonal DMs to offer, speakers will use this domain more. Examining the discrepancies observed between French and Polish, the same line of reasoning holds for the ideational domain, where thirteen different DM types are used

ideationally in Polish *versus* five DM types in French. For the rhetorical domain, however, such a relation is not found, since the two languages show a similar number of frequent DM types used in this domain (ten for Polish, eight for French). We should keep in mind here, however, that the total number of DM tokens analyzed in the two languages is very different (442 for French, 847 in Polish), making the comparison skewed because of the frequency impact. Notwithstanding this difference in the data, we would like to suggest that investigating the relation between the linguistic DM potential of a language and their domains of use is a promising avenue for further research.

As regards the expression of DM functions across the four discourse domains, we also observe a number of cross-linguistic similarities and differences. Recall that the two-dimensional annotation scheme foresees that, in principle, all DM functions can combine with all four DM domains. In the present study, we coded our data making use of eleven functions. Table 4.10 displays the domain–function mappings across the four languages studied. The domain–function mappings that are realized in all four languages are marked by the check symbol (✓), when the mapping is attested in all languages but one, the missing language is indicated preceded by the negation symbol (≠), when only one or two languages attest the domain–function mapping, these languages are mentioned.

Table 4.10 shows only two empty cells, viz. interpersonal [alternative] and interpersonal [condition], which were not expressed by a DM in any of the four languages. This is not to say that alternative or conditional DMs are never used

Table 4.10 *Cross-linguistic distribution of DM functions across discourse domains*

Function	Domain			
	Ideational	Rhetorical	Sequential	Interpersonal
Addition	✓	✓	✓	SP/PO
Alternative	✓	≠SP	≠EN	Ø
Cause	SP/PO	✓	SP/PO	FR/SP
Concession*	≠EN	✓	✓	≠EN
Condition	SP/PO	≠FR	SP/PO	Ø
Consequence	✓	✓	✓	FR/SP
Contrast*	✓	SP/EN	SP/EN	SP
Punctuation	PO	FR/PO	✓	✓
Specification	SP/PO	✓	✓	SP/PO
Temporal	≠EN	SP/PO	≠FR	SP
Topic	PO	SP/PO	✓	PO

* The function [adversative] that was used for the Polish data was recategorized here as Concession.

interpersonally in these languages, but in any case, this use would be very rare. In contrast, no fewer than sixteen cells in Table 4.10 are realized for all four languages, and an additional six cells are filled for three languages (the cells mention the language that does not attest the mapping). Thus, across the four languages, most discourse domains are compatible with several functions. In particular, the sequential domain is compatible with nearly all functions in the taxonomy across languages, whereas the interpersonal domain is the most restricted, since only two combinations are present in three or more languages, namely [punctuation] and [concession]. Cross-linguistically, it is noteworthy to observe that Spanish and Polish DMs very often express the same domain–function mappings.

Taking functions as a starting point, the two-dimensional approach also makes it possible to distinguish those functional meanings that are more tied to a specific domain and the ones that are more polyfunctional. Thus, the functions [addition], [alternative], [concession], and [consequence] are cross-linguistically realized in three of the four domains for three languages or more. Cross-linguistically, none of the functions is linked to a single domain, thus confirming the strength of the two-dimensional framework to describe the meaning and functions of DMs across languages. Of course, languages diverge in the means they provide to express these meanings and functions, balancing between functionally and semantically underdetermined DMs (DMs used in several domains and with several functions) and functionally and semantically very specific DMs (DMs used in a single domain and/or with a single function).

The results in Section 4.4 showed that the two-dimensional framework makes it possible to uncover detailed tendencies in the use of specific DMs. We believe that applying the method cross-linguistically opens up interesting perspectives in terms of DM equivalence across languages, a domain that has proven to be notoriously complex (Mortier and Degand 2009; Dupont and Zufferey, 2017) as semantic meanings and contextual use do not always align across languages.

4.6 Conclusions

This contrastive study tested the validity of a two-dimensional annotation scheme for the domains and functions of discourse markers across four spoken languages with a different typological genus, namely English (Germanic), Polish (Slavic), French, and Spanish (Romance). The analysis shows that the model can be applied to all four languages in the sample and accurately depicts the polyfunctional behavior of discourse markers as well as allows for fine-grained observations of polysemy (several functions) and polyfunctionality (several domains). Of all the theoretically possible combinations of domains and functions, only two were not attested in any language. Our comparative

findings reveal that discourse markers are primarily used to structure minor and major boundaries in the discourse flow in all languages (i.e., sequential domain) and least frequently used to relate objective facts (i.e., ideational domain). Discourse markers in all languages further share the ability to express multiple functions and domains in different contexts. The most striking difference in our dataset is the particularly high frequency of interpersonal uses of discourse markers in Spanish and Polish, which can be explained by their large number of DM types with an interpersonal core (in addition to potential methodological differences). Therefore, this study has shown that typological family cannot always account for similarities and differences in discourse-level phenomena.

This study and the annotation scheme it applies illustrate the importance of research on spoken language. The most prominent results stemming from our study concern functions of discourse markers that are specific to interactive settings (interpersonal and, to a lesser extent, sequential uses) and therefore cannot be observed in written texts. Annotation studies in written corpora are usually larger in scope and particularly useful for computational applications. However, only spoken corpora can accurately access the intricate processes of language use in interaction and how discourse is used to achieve goals and maintain interpersonal relationships. Such an endeavor is limited by the smaller size and general low availability of spoken/audio data, but the present chapter vouches for the feasibility and relevance of this line of research. This type of study also opens up cross-linguistic and cross-cultural perspectives in the prominent fields of interactional linguistics (e.g., Couper-Kuhlen and Selting 2018). Complementary approaches to DMs, combining insights from conversation analysis (e.g., Bolden 2009) and sociolinguistics (e.g., Tagliamonte 2005), would also further enrich our understanding of *how* and *why* speakers use DMs.

Further avenues of research include a detailed examination of individual differences in order to measure to what extent the observed cross-linguistic differences are idiosyncratic, as recent research has shown that individuals exhibit preferences that can relate to many variables of their background (e.g., Zufferey and Gygax 2020). This potential variation can also be checked by adding more data (more speakers) to the sample. Within the languages under scrutiny, comparison with written corpora is another promising and much wanted avenue, as speech-writing comparisons remain rare in the field (Fox Tree 2015; Kunz and Lapshinova-Koltunski 2015).

PROOF