7 Discourse Relations and Connectives across Languages and Genres

7.1 INTRODUCTION

Most languages possess a repertoire of connectives to express discourse relations. The list of connectives can even be quite extensive in some languages, mostly (but not exclusively) in the Indo-European family. For example, the French database of connectives LEXCONN lists 328 different connectives (Roze, Danlos & Muller, 2012), the Lexicon of Czech discourse connectives has 205 entries (Mírovský et al., 2017) and the Arabic Lexicon has 390 entries (Keskes, Benamara & Belguith Hadrich, 2014). Yet, comparisons between connectives expressing a similar relation have shown that most of the time, exact translation equivalents do not exist, because of frequent semantic, syntactic and register differences between them.

In this chapter, we will first present the extent of these cross-linguistic differences and discuss their implications for theories of discourse. We will make a distinction between discourse relations that seem to exist in all languages and their mapping onto specific connectives that is most of the time language specific. Cross-linguistic studies very often rely on the use of multilingual corpus data. We will therefore present the kind of data that can be used to perform contrastive studies, emphasizing their advantages and limitations.

Connectives do not only differ between languages. They are also used quite differently across different genres within the same language. These differences are particularly evident between spoken and written genres. We will also present the variations linked to genres in this chapter and underline the necessity to develop more cross-linguistic studies that are also varied in terms of genres, as such studies are still rare for the time being. Results from corpus studies comparing languages or genres have increasingly been used as input for experimental research. We will discuss in particular how observations about connective usage across genres has been important for studies analyzing discourse processing, as well as first and second language acquisition.

Taken together, we will see that analyzing variations between languages and genres is important so that results are generalizable beyond specific cases.

Comparing languages and genres raises the question of including data originally produced in one language and then translated, as translations have been said to represent a specific discourse genre, identifiable even by machine-learning algorithms (Baroni & Bernardini, 2006). We will close this chapter with a discussion of the use of connectives in translated texts and explain the importance of these observations for testing cognitive theories in the domains of discourse and translation studies.

7.2 VARIATIONS ACROSS LANGUAGES

Even though contrastive linguistics has existed as a field since the middle of the twentieth century (e.g., Lado, 1957), it has taken a new turn since the 1990s with the arrival of large multilingual corpora (Johansson, 2007). Since then, contrastive linguistics has become ever more associated with the use of corpus data. In fact, Hasselgård (2020) reports that between 1998 and 2009, 69 percent of the studies published in the journal *Languages in Contrast* were corpus-based, and this percentage increased to 83 percent for the years 2010–18. At the beginning of this section, we will therefore briefly present the methodology underlying the use of corpora to compare languages. We will then discuss studies that have taken an onomasiological perspective on cross-linguistic comparisons, starting from one or several discourse relations, and examining the various ways in which they can be signaled by connectives and other means across languages. In the last section, we will discuss studies that have taken a semasiological perspective and compared pairs of connectives across languages.

7.2.1 Methodological Aspects of Corpus-Based Contrastive Studies

Contrastive studies can be performed based on two types of corpus data: comparable and parallel corpora. Comparable corpora are simply two independent corpora produced by native speakers in monolingual contexts that are put together to identify cross-linguistic differences. In order to make such comparisons possible, the data included in these two corpora must be as similar as possible. They must minimally correspond to the same time period, be designed for a similar audience, and belong to the same genre (Johansson, 1998). Depending on the

research questions, additional variables should be controlled for comparability as well. For instance, in the case of spoken corpora, the number of speakers present in the interactions should be the same in both languages. Thus, the main challenge when using comparable corpora is that finding highly similar data in two or more languages can be quite complicated, as some genres might simply not exist in one of them, or may not be publicly available. Yet, Aijmer (2008) observes that many European languages already have a lot of available corpora, and that in many cases, existing data can simply be put together to form a comparable corpus.

Another challenge linked to the use of comparable corpora is that not only must the data be comparable, but the comparative concepts must be chosen in such a way as not to bias the comparison due to the specific labels existing or not existing in the languages to be compared, and to put maximal light on existing differences between them. Krezeszowski (1990) calls the neutral comparison platform that must be used to compare languages a *tertium comparationis*. In fact, comparing languages directly to one another based on the labels used to describe relations and connectives in one of them can often be misleading, as it may induce artificial similarities or differences between them. Let us take an example. If English and French are compared based on the presence of a discourse relation called "chosen alternative" in an annotated dataset of connectives, these languages would appear to be totally different, as only English would have such a label linked to the uses of the connective *instead*. This is because French does not have a lexicalized connective similar to English to express this relation. However, alternative lexicalizations exist, such as *à la place* or *plutôt* that can perform the same function in some contexts. Conversely, if French and English are compared based on the existence of connectives expressing a causal relation, the two languages will misleadingly appear to be identical, whereas the lexicon of causal connectives in both languages exhibit important differences, as we will see below. In sum, it is crucial to find a tertium comparationis that involves comparable and language-independent concepts (König, 2012) with an adequate degree of granularity to reveal cross-linguistic differences. This can be a major challenge, especially since the very existence of universal categories has been put into question by some authors (e.g., Evans & Levinson, 2009).

Finally, in order to be compared, connectives or discourse relations found in comparable corpora must be annotated in both languages. Yet, such annotations can be quite difficult to perform. For instance, Cartoni, Zufferey and Meyer (2013a) trained two annotators to

categorize occurrences of the English connective *while* with four different labels corresponding to its different meanings in the Penn Discourse Treebank (PDTB) (see Chapter 2). They report that even after several rounds of training, the two annotators did not reach a high level of agreement, as only 68 percent of the annotations were congruent. This corresponds to a kappa value of 0.43, which does not reflect a reliable level of agreement (Arstein & Poesio, 2008). In a similar vein, Spooren and Degand (2010) discuss the difficulty of annotating coherence relations with a high degree of reliability. They advocate the use of very precise and transparent coding schemes, so that biases are at least clearly documented, and can be discussed in subsequent research.

An alternative solution to the use of comparable corpora, which solves many of its problems, is to resort to the use of parallel corpora, in other words, corpora containing original texts in one language and their translation in another language. The main advantage of parallel corpora over comparable ones is that they provide a direct way to compare languages, through the observation of translations. Contrary to sense annotations, this method, called *translation spotting* (Cartoni, Zufferey & Meyer, 2013b), is very easy and reliable, and it can even partly be performed automatically (Véronis & Langlais, 2000). But parallel corpora also have a number of shortcomings. The major one is that these corpora are still quite limited. In fact, in many genres, translations are simply not produced, which limits their availability to specific cases such as literary texts, newspaper articles, subtitles, and text produced by multilingual organizations and countries. In addition, these corpora are available for a limited set of language pairs only. Another problem with these corpora is that comparing original texts to translations can represent a bias in itself, as translated language represents a specific genre with its own characteristics (see Section 7.4).

In sum, both comparable and parallel corpora have their own problems and advantages. Importantly, these problems are different in both cases, and they can be at least partly overcome by using both types of data simultaneously. In this respect, the creation of large multilingual corpora such as the Europarl corpus[1] (Koehn, 2005), consisting of debates at the European Parliament, have provided a major step ahead. The principle of equality between all languages of the European Union means that every person speaks in their own language, and the speeches are later translated into all the other languages (24 at the time of writing). This means that portions of the Europarl corpus can

[1] www.statmt.org/europarl/.

be assembled to perform various types of comparisons. First and foremost, it allows for the creation of bidirectional corpora, in which both languages are alternatively sourced and targeted, thus avoiding the bias of looking at only one translation direction. It can also be used as a comparable corpus to compare original language productions in highly similar contexts, as all deputies speak about similar topics, in a similar environment, and samples can be taken from the same time period. Finally, it can be used to compare translations performed from various source languages, in order to investigate their influence on translated texts (see Section 7.4). Thanks to the simultaneous availability of parallel and comparable corpora, sense annotations can be combined with translation spotting, which provides a fuller picture of cross-linguistic equivalences and differences. In this chapter, we will present several studies performed on bidirectional corpora, many of them using Europarl.

7.2.2 Discourse Relations across Languages

In most models of discourse, coherence is defined as a cognitive notion (see Chapters 2 and 6). It can therefore be expected that similar relations linking discourse segments will be found across languages. There is indeed some evidence that different languages use a similar set of relations. In fact, most major models of discourse (see Chapter 2) have been used to annotate corpus data in several languages. For example, the PDTB framework has been used to annotate relations in Arabic, Chinese, Czech, Danish, Dutch, Hindi and Turkish. Rhetorical Structure Theory (RST) has been used to annotate data in Basque, Dutch, German, English, Portuguese and Spanish, and corpora annotated with Segmented Discourse Representation Theory (SDRT) exist in Arabic, French and English (Benamara Zitoune & Taboada, 2015). In some cases, the taxonomies of relations had to be modified, with the addition or elimination of some relations, or even an elimination of the hierarchical nature of some models like the PDTB (Prasad, Webber & Joshi, 2014). But the major cross-linguistic differences come from the mappings between connectives and discourse relations, rather than the existence of similar discourse relations.

Another line of evidence indicating that similar discourse relations hold across languages comes from the multilingual annotation experiment conducted by Zufferey and Degand (2017). They used the PDTB framework to annotate discourse relations in five Indo-European languages (Dutch, English, French, German and Italian). They also found that the framework could be used in all of them with only minimal changes. For instance, they added a relation of parallelism

to the comparison section, but this is not only due to the existence of specific connectives in some languages, but rather to a relation that appears to be missing in English as well, as the meaning of connectives like *similarly* did not find a straightforward tag in the original PDTB scheme. In the same vein, Kolachina et al. (2012) also suggested the addition of a "similarity" tag in the comparison section in the revised version used to annotate Hindi. One of the most important results from Zufferey and Degand's (2017) study on multilingual annotation is the observation that the meaning of some connectives requires several sense tags to be accounted for rather than new tags. For example, the French connective *tant que* simultaneously conveys a meaning of temporality and condition in all of its uses. Allowing the annotation of relations with double tags therefore represents a necessary adjustment for these connectives.

In another line of research, some studies have focused on one specific relation and observed the various ways in which it can be signaled across languages. This approach also relies on an onomasiological perspective, as it starts from the relation rather than from specific connectives. One study that has taken this perspective was conducted by Taboada and de los Ángeles Gómez-González (2012), who focused on the relation of concession in English and Spanish, a notion that they consider to be similar across languages. They annotated comparable corpora from the written and the spoken modes (see Section 7.3.1) in both languages using the RST framework (see Chapter 2). Yet, they did also include lexical constraints on their analysis, as they extracted for annotation all the relations that were signalled by a marker, a notion broadly defined to include connectives, paraphrases and alternative lexicalizations. This approach enabled them to find concessive relations automatically, but it excluded relations that were conveyed implicitly. Based on their annotations, they conclude that the relation of concession functions in a very similar manner in English and Spanish, as most of the differences they found were linked to the different genres compared in the study (see Section 7.3).

In another study, Cuenca, Postolea and Visconti (2019) compared the signaling of contrastive relations in Spanish, Catalan, Romanian and Italian. They used a corpus-informed rather than a corpus-based approach, in so far as their study focuses on published materials in all languages rather than make direct comparisons between them using comparable or parallel corpus data. They also include a historical component to their analysis, by comparing cognate connectives, in other words connectives historically coming from the same source, such as *al contrario* (Spanish), *al contrari* (Catalan), *al contrario* (Italian)

and *din contra* (Romanian). All of these connectives have their source in the word for 'contrary' in the different languages. The main conclusion from this analysis is that cognate connectives often diverge in modern Romance languages, such as the connectives *pero* (Spanish), *però* (Catalan) and *però* (Italian). In Italian, *però* is only occasionally used to mark nonexclusive contrast, contrary to Spanish and Catalan, because a concurrent form (*ma*) exists in this language but not in the other two. This study is interesting because it underlines the fact that cognate connectives often evolve in such a way as to make them more different from one another. However, as it does not include corpus-based comparisons made on parallel data, the actual translation equivalents between connectives cannot be inferred based on this research. For this, a more fine-grained comparison between pairs of connectives is necessary.

7.2.3 Discourse Connectives across Languages

While studies focusing on discourse relations have found that the same relations exist and are expressed by similar means across languages, studies focusing on specific connectives have almost always revealed important differences between them. One of the reasons for these differences is that connectives are often polyfunctional (see Chapter 3), and equivalents are often found for just one of their meanings. For instance, the French connective *en effet* can express either a relation of cause in the sentence-initial position, or a relation of confirmation (similar to the English *indeed*) in the clause-medial and clause-final position (Charolles & Fagard, 2012). While its confirmative uses have translation equivalents in other languages like German and English, its causal uses do not, as illustrated by the fact that these relations are mostly left implicit in translations (Zufferey, 2016). Similarly, a contrastive study comparing French *en effet* and Russian *v samon dele* (Iordanskaja & Mel'čuk, 1999) also found that the meaning of these two connectives is only partially related, as both have functions that are better translated by other connectives or paraphrases. To take another example, in Lithuanian, one of the two most frequent contrastive connectives in spoken language, the connective *o*, does not seem to have a translation equivalent in English (Šliogerienė, Valūnaitė Oleškevičienė & Asijavičiūtė, 2015).

Another reason why connectives often don't have translation equivalents is that they all come with their own syntactic constraints, that are often not shared across languages. For instance, the English connective *also* can be used to convey additive relations in all syntactic positions, whereas the French additive connective *aussi* can only convey

this meaning in the clause-medial or final position, as in the clause-initial position it takes a meaning of consequence (Roze, Danlos & Muller, 2012). Finally, there are also differences in terms of register between connectives expressing a similar meaning. In French, the causal connective *car* is now mostly used in the written mode (Simon & Degand, 2007) and perceived as formal by French speakers (Zufferey, 2012), but its Dutch counterpart *want* does not have similar restrictions (Spooren et al., 2010). In the rest of this section, we present studies that illustrate these differences using corpus-based methods.

Gast (2019) compared three concessive connectives: the English *although*, the German *obwohl* and the Spanish *aunque* using comparable data from the Europarl corpus. The author annotated many different dimensions of connective usage, such as the type of concession involved, the structural properties of the concessive clause, and the level of linking (propositional, illocutionary or textual). His results demonstrate that each connective has its own specificities not shared by the other two. First, the German *obwohl* seems to be more restricted than the other two in the type of concessions that it marks, and it is not used in preposed concessive relations, contrary to *although*, which is often used in this syntactic position. The author explains the differences between *obwohl* and the other connectives by the fact that German uses a specific word order for subordinate clauses (verb final position) and that this syntactic pattern blocks further distributional extensions for this connective (see Chapter 4). This study is very informative because of the number of dimensions that are compared. However, it does not make use of the parallel data also provided by Europarl, so that comparisons between the annotated features and the translation equivalents cannot be made.

In the domain of causal relations, Pit (2007) provides a comparison between four causal connectives in Dutch (*doordat*, *omdat*, *want*, *aangezien*), three in French (*car*, *parce que*, *puisque*) and three in German (*weil*, *denn*, *da*) based on comparable corpora made of newspaper articles and narrative texts. The main comparison criterion in this study is the degree of subjectivity of the causal relation (see Chapters 3 and 6 for a discussion of the distinction between objective and subjective causality). She reports that in all three languages, there is a scale on the degree of subjectivity that each connective typically conveys. The Dutch connective *doordat* is the most strongly associated with objective contexts. Then come the connectives *omdat* in Dutch, *weil* in German and *parce que* in French, which are slightly less strongly objective, but more importantly have a similar (and rather low) degree of subjectivity on the scale. In contrast, the connectives *want* and *aangezien* in Dutch, *denn* in

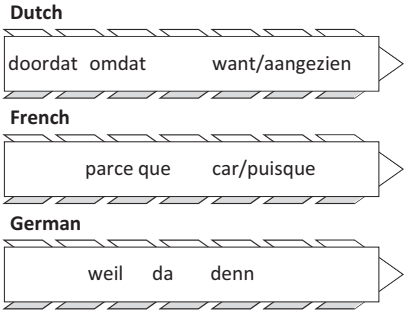

Figure 7.1 Scale of subjectivity across languages (adapted from Pit, 2007)

German and *car* and *puisque* in French have a high degree of subjectivity that is comparable across languages. Finally, the German connective *da* is situated between *weil* and *denn* on the subjectivity scale, and does not have counterparts in other languages, as illustrated in Figure 7.1.

In another corpus study involving the use of comparable and parallel corpora, Degand (2004) more specifically compared the Dutch connective *aangezien* to the French *puisque*, as both are often presented as a translation pair in bilingual dictionaries. Yet, Degand found that this is far from being the case, especially in the Dutch–French direction, as *aangezien* is translated by *puisque* in only 8 percent of the occurrences, against 48 percent of translations of *puisque* by *aangezien* in the French–Dutch direction. This study thus revealed that equivalences can be quite different in both translation directions. Degand attributes this discrepancy to the fact that *puisque* appears to be more strongly subjective than *aangezien*. For this reason, less strongly subjective connectives such as *étant donné que* are better translation equivalents for *aangezien*. In the other direction, as Dutch does not have an equally strongly subjective connective, *aangezien* is used to translate *puisque*, even though their meanings do not fully coincide. This study illustrates the advantage of combining corpus annotations in comparable data with translation data.

Using a similar method, Zufferey and Cartoni (2012) compared the meanings and translations of causal connectives in French (*car*, *parce que*, *puisque*) and English (*because*, *since*, *as*). They annotated 200 occurrences of each connective in original data from the Europarl corpus, using two dimensions. The first was whether the connective conveyed an objective or a subjective relation and the second was whether the cause segment introduced by the connective conveyed given or new

information. The rationale for this second dimension comes from previous studies on French connectives, arguing that the main difference between the subjective connectives *car* and *puisque* is that *car* introduces new information whereas *puisque* introduces information that is already known or at least highly obvious for the hearer (Groupe lambda-l, 1975; Franken, 1996; Zufferey, 2014). In her analysis of French causal connectives, Pit (2007: 73) had already compared *car* and *puisque* in terms of givenness in corpus data, and found that this difference is borne out. However, she did not extend it to the other languages in her study, and her sample of occurrences was rather limited (50 for each connective).

In their analysis, Zufferey and Cartoni found that none of the causal connectives in English and French form an exact translation pair, and their different semantic profiles can be used to predict the chosen translation equivalents. For example, the English *because* is used to convey both objective and subjective relations. When it is used to convey an objective relation, its main translation equivalent in French is *parce que*, but when it conveys a subjective relation, its main translation equivalent is *car*. Another example is the connective *since*, mostly translated by *car* when it conveys new information, and by *puisque* and *étant donné que* when it conveys given information. Their results also confirm that *puisque* does not have an exact translation equivalent in English. As a result, this connective is used five times more in original French texts compared to translations.

Zufferey (to appear) made a similar comparison between French and Spanish causal connectives (*porque, ya que, puesto que*). Like English, Spanish causal connectives do not differ systematically in terms of their degree of subjectivity in monolingual corpus studies (Santana et al., 2018). In contrast, Zufferey found that these connectives differ in terms of their propensity to convey new or given information. Using the Europarl corpus, she found that *porque* is mainly used to convey new information, *ya que* also conveys a majority of new information but can also be used for given information in some contexts, and *puesto que* is used mostly for given information. Spanish and French causal connectives were also found to differ. The connective *porque* is used more often than *parce que* to convey subjective relations, and again the translation equivalent in French varies depending on this factor (*parce que* for objective relations and *car* for subjective ones). The Spanish *ya que* is more often used to convey given relations than the French *car*. In turn, *puisque* is more strongly subjective than *puesto que*.

In sum, all the contrastive studies presented in this section reveal important cross-linguistic differences between connectives, even

between closely related languages. The combination of comparable and parallel corpus data has in addition revealed that the semantic profile of connectives can be related to the chosen equivalent in translations.

7.3 VARIATIONS ACROSS GENRES

In addition to analyzing the way connectives vary across languages, studies have also investigated the impact of genre for variations in the use of connectives. In this section, we review studies that have compared the use of discourse relations and connectives across various spoken and written genres. We will see that the communication of various discourse relations is indeed sensitive to genre. These variations are more specifically linked to the different connectives used in all of them, as well as the registers they belong to. At the end of this section, we discuss the impact of genre for the processing and acquisition of discourse relations and connectives.

7.3.1 Discourse Relations and Connectives across Spoken and Written Genres

In Chapter 2, we introduced different models used to annotate discourse relations, and we discussed their applicability to different languages in this chapter. While we concluded that these models can usually be applied to different languages with minimal changes, it is not clear that they would be equally successful for annotating discourse relations in all genres. In fact, several studies have questioned the fact that the models of discourse annotation, designed mostly for use in written language genres, can equally serve to annotate spoken data. For this reason, an alternative model, also encompassing tags accounting for typically spoken uses of connectives, has been suggested (Crible & Degand, 2019a). This model includes for instance the functions of quoting, topic resuming, disagreeing, etc. that are typical of spoken language interactions. In preliminary annotation experiments, it was found to be suitable to annotate connectives in written and spoken data, in both English and French (Crible & Zufferey, 2015). In addition to the existence of more interactive genres in the spoken mode, the justification for using a different kind of annotation framework for spoken data comes from the fact that in this mode, a limited number of connectives are used with a very high frequency, and with functions that they do not fulfill in written genres.

To make a case in point, in spoken data, the English connective *and* can take as many as eleven different functions, against only four in writing (Crible & Cuenca, 2017: 159). These functions include topic-shift, specification, contrast and temporality, to name but a few. The reason for this versatility is that *and* is an underspecified connective whose meaning can be enriched in context (Blakemore & Carston, 2005). In order to assess the role of genre for the use of underspecified connectives, Crible and Demberg (2020) analyzed the uses of *and* across three types of spoken data involving a different register: informal (face-to-face and phone conversations), semi-formal (interviews, classroom lessons) and formal (news broadcasts, political speeches). Their hypothesis was that some functions of *and* would be bound to specific registers. More specifically, they hypothesized that the use of *and* to mark contrastive relations (a non-canonical use of this connective that has a core meaning of addition) would only be found in informal registers, as more specific connectives would be chosen in more formal registers. In other words, in formal registers, they expected *and* to be closer to its core meaning of addition. Using a multi-genre corpus (Crible, 2018), they extracted all the uses of *and*, as well as two other underspecified connectives *so* and *but* in all genres. Their results indicate that *and* is indeed used more broadly in informal contexts. Similarly, *but* and *so* are also more used with their core meaning (contrast and consequence) in formal contexts. Thus, the relation between the use of underspecified connectives to signal a discourse relation and the different registers linked to various genres is confirmed in this corpus study.

These observations raise the question of determining whether the semantic profile of connectives established in corpus data from genres belonging to the written mode remain valid in spoken data. One study that addressed this issue was conducted by Simon and Degand (2007) who compared the semantic and prosodic profile of the French connectives *car* and *parce que*. The semantic comparison relied on the notion of subjectivity. They found an important difference between the two modes. While in writing, *parce que* appeared to be a predominantly objective connective and *car* was more subjective, in speech *parce que* was equally used for objective and subjective relations. This difference is due to the fact that in contemporary spoken French, *car* is very seldom used. In fact, they report that its frequency drops from 0.4 percent to 0.02 percent from the written to the spoken mode, while that of *parce que* increases from 0.32 percent to 3.7 percent. Degand and Fagard (2012) argue that this imbalance between the written and the spoken modes illustrates a situation of change in progress, in which one connective (*parce que*) is in the process of replacing another one

(*car*), and these changes always start in spoken communication before spreading to written genres.

In contrast, Spooren et al. (2010) found that the two causal connectives *omdat* and *want* in Dutch do not exhibit a similar difference between written and spoken genres, as *want* remains more subjective than *omdat* in spoken data. Sanders and Spooren (2015) also report that *want* remains more strongly associated to the communication of subjective relations than *omdat* across three different genres: written texts, conversations and chat interactions. Yet each connective has a mode of choice: *want* was the most frequent connective in spoken data whereas *omdat* was the most frequent one in written data. In that sense, mode also influences the use of Dutch causal connectives.

Similarly, Li, Sanders and Evers-Vermeul (2016) found that three causal connectives from Mandarin Chinese (*jiran, yinwei, youyu*) have a robust profile across three different written genres: news reports, novels and opinion pieces. In all genres, *jiran* expresses very subjective relations, whereas the connective *youyu* specializes in the communication of objective relations, and *yinwei* is situated between the other two connectives. In another study on Mandarin Chinese connectives indicating relations of consequence, Li, Evers-Vermeul and Sanders (2013) also found that the profile of several of them was robust across genres. There were two exceptions though: the connectives *suoyi* and *yinci* that resemble the English *so* and *therefore* were variable across genres. While these connectives were used more to convey subjective relations in news reports and opinion pieces, they conveyed more objective relations in novels. This indicates that genre interacts with the meaning of some but not all connectives. The authors argue that the differences between them may be related to the strength of their core meaning, as vaguer connectives may be more variable across genres, but this hypothesis still needs to be assessed in future research.

In a study on English, Andersson and Sunberg (2022) analyzed four connectives used to convey cause–result relations (*so, therefore, as a result, for this reason*) in various genres typically involving language with various register levels, from the spoken and the written modes. They found that contrary to Dutch or Mandarin Chinese, connectives are not systematically associated with a certain degree of subjectivity in English. The connective *so*, especially, is variable across domains, contrary to *therefore* that is always used to communicate epistemic (i.e., subjective) relations. In addition, *so* is the most likely connective to convey consequence relations in English, independently of the degree of subjectivity of the relation, which indicates again that subjectivity is not a strong factor to categorize English connectives. There were also clear register effects in

the data, as some connectives like *therefore* were used more in contexts involving a high register, both in speech and in writing, whereas *so* was linked to the use of less formal registers. The connectives *as a result* and *for this reason* had a very low frequency in spoken language. It seems therefore that register is also an important factor to account for the use of connectives in some genres. We will see below that register is also an important factor for language processing and acquisition.

All the studies reviewed so far have underlined differences linked to genres pertaining to the spoken mode, as opposed to the written mode. Yet, other studies have also found differences between genres pertaining to the written mode. For example, Smith and Frawley (1983) compared the use of connectives in English across four written genres: fiction, journalism, religion and science. They noted different uses of discourse relations and connectives in all of them. For instance, they found a higher number of adversative relations in fiction compared to the other genres, especially science, that had very few such relations. Similarly, temporal relations were more prevalent in journalism compared to the other genres. Differences were also found at the level of individual connectives. For example, *yet* and *although* were used much more often in religious texts compared to *but*, a connective used more frequently in the other genres. Similarly, the causal connective *for* was used mostly in the religious genre and to a lesser extent in fictional texts. This difference can be linked to the high register associated to this connective in modern English. Liu (2008) also compared the use of connectives across five genres in the British National Corpus (BNC corpus), and found that the number of connectives used was quite variable across genres. Connectives were most frequent as a whole in academic texts, and least frequent in the news genre. The type of connectives used also varied across genres. For example, connectives indicating temporal simultaneity such as *meanwhile* and *in the meantime* were mostly found in the news genre.

In conclusion, all the studies reported in this section indicate that genre matters for connective usage, even though few of them have made comparisons between genres belonging to the same mode. We now look at the effects of these differences in the way discourse relations conveyed by connectives in different genres are processed and acquired in first and second language.

7.3.2 The Impact of Genre for Language Processing and Acquisition

In Chapter 6 (see Section 6.6), we discussed the impact of cross-linguistic differences for discourse processing. We saw, for example, that the way

a discourse relation is encoded in a language can have an impact on processing with the example of *want* in Dutch, a connective that involved an immediate instruction of subjectivity causing readers to slow down (Canestrelli, Mak & Sanders, 2013). Another question is whether reading relations belonging to a given genre can also have an impact on discourse processing. Canestrelli, Mak and Sanders (2016) investigated this issue in an eye-tracking experiment in which readers had to read subjective relations with *want* appearing in short texts reproducing the features of two different genres: news items and letters to the editor. The authors hypothesized that letters to the editor would involve a more subjective style overall compared to news items. As a result, they predicted that reading these texts would take more time, as previous research has shown that inferring subjectivity is a costly process (see Chapter 6). However, they also thought that encountering *want* in highly subjective contexts would involve a smaller delay or even no delay at all compared to objective contexts, as readers would already be oriented towards the subjective domain. The first hypothesis was borne out, as reading subjective texts indeed took longer, but the connective *want* involved the same processing delay across both genres. In that sense, genre did not have an impact on processing, or at least did not prevent the processing cost associated with subjective connectives.

In another experiment in French, Zufferey et al. (2018) compared the processing of objective and subjective causal relations conveyed by *car* and *parce que*. Recall that in French, the connective *car* is seldom used in the spoken mode, and is associated with a formal register by French speakers. The authors hypothesized that because *car* is not used anymore in speech, French readers may not have strong intuitions about its meaning, and therefore not use it as a clue to infer subjectivity. They found that French readers slow down when they encounter *car*, but they do so independently of the type of relation it conveys (objective or subjective). The authors conclude the presence of a register effect: having to process a less familiar connective belonging to a formal register induces a delay compared to connectives frequently used in both written and spoken genres like *parce que*.

The role of genre, and more specifically the formal registers associated to some written genres, found further confirmation in a study by Wetzel, Zufferey and Gygax (2022), who compared reading times between correct and incorrect causal and concessive relations. More specifically, they compared the processing disruptions created by inappropriate connectives with different profiles in terms of semantics and register: in a first experiment they used connectives frequently used in spoken language (*mais* and *alors*), in a second experiment they used monofunctional connectives mostly used in the written mode

(*néanmoins* and *ainsi*) and in a third experiment they used polyfunctional connectives bound to the written mode that can be used to convey different relations depending on context (*or* and *aussi*). They found that even if native speakers' ability to detect inappropriate uses of connectives was robust across all experiments, their sensitivity decreased with connectives from the written mode. These experiments therefore provide further confirmation for the role of register on speakers' use of connectives during discourse processing.

The greater complexity of connectives bound to written genres was also found in studies targeting first language acquisition during teenage years. For example, Nippold, Schwarz and Undlin (1992) tested the comprehension of connectives from the written mode by teenagers and young adults, and found that connectives from the written mode are not fully acquired in this age group (see Chapter 8). Similarly, Tskhovrebova, Zufferey and Gygax (2022) tested the ability of French-speaking teenagers to understand four connectives from the written mode, and found that teenagers still perform poorer than adults even during high school. These difficulties are not generalized, however, as teenagers have a better performance (but not at the level of adults) when the task is made simple enough and the connectives are monofunctional (Tskhovrebova, Zufferey & Tribushinina, 2022). We discuss this issue in more detail in Chapter 8.

In the context of second language learning, connectives belonging to written genres are also an area of difficulty for learners. However, the factor of genre is not always the most relevant one to explain the problems encountered by this population. Register seems to matter as well, as learners master better connectives belonging to an informal than a formal register (Wetzel, Zufferey & Gygax, 2020). We will come back to this issue in Chapter 9.

7.4 STUDIES COMBINING VARIATIONS ACROSS LANGUAGES AND GENRES

One of the main limitations of current contrastive studies is that in many of them, languages are treated as uniform entities, and variations between genres and registers are not considered. Yet, linguists performing contrastive analyses are increasingly recognizing the need to extend cross-linguistic studies in order to account for genre variations (e.g., Johansson, 2007). In several areas of cross-linguistic research, the inclusion of genre variations has led to fruitful results. For instance, Granger (2014) found that genres have an impact on the use of lexical bundles in French and English. In the domain of

connectives, studies including both cross-linguistic and cross-genre comparisons are still few and far between. We review them in this section.

First, the study mentioned above on concessive connectives in English and Spanish (Taboada & de los Ángeles Gómez González, 2012) also compared the use of connectives across a spoken corpus (telephone conversations) and a written corpus (online book and movie reviews) in the two languages. The authors found that language users make a different use of concessions in the two modes. In speech, concessions serve most of all to correct misunderstandings, and in the written mode they serve to qualify the writer's own opinions. In line with Crible and Cuenca (2017), they also report that concessions are conveyed with a lesser variety of connectives in spoken language, as *but* and *pero* are predominantly used whereas in writing the array of markers is much broader. Overall, they report that differences between genres are more numerous and prevalent than differences between languages. In her study on the use of discourse markers (see Chapter 3) across eight genres in English and French, Crible (2018) also found that genres matter more than language for the variations observed.

Kunz and Lapshinova-Koltunski (2015) analyzed the use of connectives and other cohesion markers (see Chapter 1) in a corpus of English and German containing ten different registers in each language, eight from the written mode and two from the spoken mode. At a general level, they found differences between the two languages in their use of cohesion markers. Regarding the use of connectives more specifically, they report that German uses more explicit connectives than English, thus confirming other observations about the low level of connectivity in English (Vinay & Darbelnet, 1995). They also report that German shows more variation overall in the use of cohesive devices across registers, but it is difficult to infer more specific differences about specific discourse relations from their data analyses.

Finally, Dupont and Zufferey (2017) investigated the use of four concessive connectives in English (*however, nevertheless, nonetheless, yet*) and four French concessive connectives (*toutefois, néanmoins, cependant, pourtant*) in three parallel corpora: the Europarl corpus of parliamentary debates, a selection of texts from the newspaper section of the PLECI corpus,[2] and the TED Talks corpus[3] (Cettolo Girardi & Federico,

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/pleci.html.
[3] https://yohasebe.com/tcse/.

2012). In their analysis, they considered three dimensions of variation: the translation direction, the genre, and the level of expertise of translators. They report that all three dimensions matter for the observed translations of concessive connectives. First, the type of translation equivalent was often divergent between the two translation directions. But importantly, these differences were dependent on genre. For example, the English connective *nevertheless* is translated by *néanmoins* in 68 percent of the occurrences in the TED corpus, against only 39 percent in Europarl. More generally, some genres like parliamentary debates seem to encourage more specific translations, whereas the news genre has more generic translations with connectives like *but* and *mais*, and also more cases of implicit translations. Conversely, the frequency of each connective differs across genres, and these differences are more pronounced in English.

It is clear from the studies reported in this section that cross-linguistic studies of connectives should in the future strive to include more genre diversity, as both variables seem to have an impact on the use of discourse relations and connectives. The corpus studies reported in this section illustrate the complex intricacies between these two factors.

7.5 CONNECTIVES IN TRANSLATIONS

Discourse connectives are well known to be volatile in translations, as translators often add and remove them, or translate them by other lexical or syntactic means (Halverson, 2004; Denturck, 2012). These changes reflect the fact that connectives often don't have exact translation equivalents, as we discussed above. While being problematic for translators, connectives represent a valuable area of study for scholars interested in translation theory. Indeed, it has been suggested that translations represent a specific textual genre, distinct from texts originally produced in a language. The specificities of translations have been linked to the existence of universals of translation (Mauranen & Kujamäki, 2004), in other words, specificities emerging from the translation process itself. These universals include among other things a tendency for translated texts to be lexically and structurally simpler than original texts, to under-represent items that are specific to the target language, and to be more explicit than original texts due to a greater use of cohesion markers (e.g., Laviosa, 2009).

Connectives therefore represent a very interesting case study to assess the existence of a translation universal of explicitation. This

hypothesis has been tested in a number of studies, but many of them could not lead to strong conclusions because they were limited to one single language pair, and the analyses they provided were qualitative rather than quantitative. Taking advantage of the big multilingual corpus Europarl, Zufferey and Cartoni (2014) assessed the existence of a universal tendency for explicitation by comparing the use of causal connectives in French and English texts translated from four different source languages (Italian, German, Spanish, French/English). They counted the number of times four French causal connectives (*parce que, car, puisque, étant donné que*) and three English connectives (*because, since, given that*) were added in translations. They considered that explicitation had taken place when the relation was fully implicit in the source language, in other words, the segments were linked only by a comma or a full stop, and a causal connective was present in the translation. Their results indicate that explicitation does indeed seem to be a universal tendency in translations, in that the proportion of added connectives was similar for all source languages, in both target languages. However, they found significant differences between connectives. Connectives like *parce que* and *because* were almost never used for explicitation, whereas connectives like *puisque* and *given that* were very frequently used for explicitations. They relate this difference to the different semantic profile of connectives, as the connectives used for explicitation were highly subjective and marked a relation as given (see Section 7.2). It seems therefore that translators feel the need to mark subjectivity, and to signal givenness to their audience. In sum, the existence of explicitation as a universal of translation was reinforced based on this study, but one of its weaknesses is that it is limited to closely related languages. In another study on English and Chinese, Xiao and Dai (2014) also found evidence of explicitation phenomena, thus strengthening the hypothesis further.

In addition to explicitations, looking at the cases when connectives are removed in translations, in other words implicitations, is also very interesting from the point of view of discourse. We argued in Chapter 6 that there is an important cognitive difference between continuous relations that can easily be reconstructed by inference, like causal and additive relations, and discontinuous relations that need to be marked explicitly like concessive relations. If this difference does indeed represent a cognitive constraint, it should have an impact cross-linguistically, and prevent translators from removing connectives from discontinuous relations that are not signaled by alternative means. Hoek et al. (2017) investigated cases where connectives encoding several types of coherence relations were removed in translated

texts, and found that implicitation is linked to the degree of cognitive complexity of the relation, as defined in the Cognitive approach to Coherence Relation (CCR) framework (see Chapter 2). Cognitively complex relations like concessions are indeed left implicit less often than simpler relations like causal relations. Zufferey (2016) looked at implicit translations conveyed by three polyfunctional French connectives (*or, en effet, dans la mesure où*) that can each convey a continuous or a discontinuous relation. Results indicate that for all connectives and in the three target languages (German, English, Spanish), connectives were significantly more frequently left implicit in translations when they conveyed a continuous relation than a discontinuous one, thus providing further confirmation for the validity of the cognitive constraint of continuity.

In a parallel corpus study of Lithuanian and English connectives based on the TED talk corpus, Valūnaitė Oleškevičienė et al. (2022) confirmed the results presented so far: most of the explicit discourse relations in English were translated explicitly in Lithuanian (about 80 percent of the occurrences) and the remaining 20 percent corresponded to a large extent to uses of *and* in English source texts, hence a continuous additive relation. Analyzing academic papers published in Catalan and translated into English, Cuenca (2022) also found a role of the type of discourse relation, but also of syntactic factors and other specific criteria related to the language pair in question.

In sum, connective usage varies in translations, but these variations are not random: explicitation occurs mostly for continuous relations that were often left implicit in source texts such as causal ones, and conversely these relations are also the ones often left implicit in translated data. In contrast, leaving discontinuous relations implicit creates difficulties that are recurrent across languages and that limit translators' choices.

7.6 SUMMARY

This chapter presented variations in connective usage across languages, and across different genres and registers within the same language. One of the main conclusions from cross-linguistic studies is that similar discourse relations exist across languages, as a way to create coherence within text and discourse. The encoding of these relations in specific connectives is, in striking contrast, almost always language specific. Connectives do indeed differ between languages in the number and type of meanings they convey, their syntactic restrictions,

and the genres and registers in which they are typically used. These differences have been repeatedly observed, even between closely related languages such as French and Spanish, or German and Dutch. In the second part of the chapter, we observed that variations are also wide ranging between genres, especially between those belonging to the written and to the spoken modes, but also within one mode. These differences have an impact of language processing, as even native speakers have less clear-cut intuitions about connectives bound to the written mode, and these connectives are also acquired later by teenagers. Cross-linguistic differences observed in parallel directional corpora can also be used to study connectives in translated texts. These observations are crucial to test different theories in translation and discourse studies, such as the existence of a translation universal of explicitation, and the role of continuity for the implicit communication of discourse relations.

DISCUSSION POINTS

- In what respects do connectives differ between languages?
- Why are connectives difficult to translate?
- Why do you think that so many connectives are polyfunctional, and why are all these meanings almost never fully convergent between languages? (Hint: look back at the discussion on the evolution of connectives in diachrony in Chapter 5.)

FURTHER READING

Johansson (2007) is a must read to get acquainted with the use of multilingual corpora. A cross-linguistic study of connectives and other discourse markers in eight different genres can be found in the study by Crible (2018). Good examples of the use of parallel corpora to study specific connectives across languages and genres are Dupont and Zufferey (2017), and Degand (2004). The notion of genre and its application to discourse is discussed in detail in the book edited by Stukker, Spooren and Steen (2016). The notion of translation universals is presented in Mauranen and Kujamäki (2004) and discussed more succinctly by Laviosa (2009).