UCLouvain

FIAL
Un éventail de formations, une multitude d'horizons

Faculté de philosophie, arts et lettres

Discourse markers in reality TV and film dialogues

A contrastive study of *you know, so, well* and their French and Dutch translations in subtitles

Auteur : Benoît Lefin
Promoteur(s) : Prof. Liesbeth Degand
Année académique 2022-2023
Master [60] en langues et lettres modernes, orientation germaniques

Faculté de philosophie, arts et lettres

Discourse markers in reality TV and film dialogues

A contrastive study of *you know, so, well* and their French and Dutch translations in subtitles

Auteur : Benoît Lefin
Promoteur(s) : Prof. Liesbeth Degand
Année académique 2022-2023
Master [60] en langues et lettres modernes, orientation germaniques

Acknowledgments

I would first like to thank my thesis supervisor Prof. Liesbeth Degand for her guidance, support, and precious comments throughout my master programme. Her availability and insight have been invaluable and have played a crucial role in the success of this thesis.

I would also like to thank all the professors of UCLouvain who directly or indirectly participate in the development of my thesis. The journey of my master programme, whose thesis is the culmination, has been an unforgettable experience thanks to them.

Lastly, I would like to thank my family and friends for the unwavering trust and support.

Table of contents

List of figures

List of tables

List of abbreviations

Abbreviation	Definition
ADD	Addition
AGR	Agreeing
ALT	Alternative
CCS	Concession
CND	Condition
CSQ	Consequence
CSV	Comma separated Variables
DIS	Disagreeing
DM	Discourse marker
EFL	English as foreign language
FR	French
HDG	Hedging
IDE	Ideational
INT	Interpersonal
L2	Second language
LiB	*Love is Blind*
NL	Dutch (Nederlands)
PDTB	Penn Discourse Treebank
QUO	Quoting
RHE	Rhetorical
SEQ	Sequential
SPE	Specification
TED	Technology, Entertainment, Design
TMP	Temporal
TOP	Topic
XML	Extensible Markup Language

1 Introduction

Speech is the origin of language. Oral comes before writing and even if both have unique characteristics, speech remains central to understanding how humans communicate. Some specific devices massively used in oral conversations are the discourse markers. Schiffrin (1987) is the first to identify and name these devices as discourse markers, and they have been the subject of many studies. There are still debates about what they should be called and what they could define (Stubbe & Holmes, 1995; Fraser, 1996; Rouchota, 1996; Aijmer & Simon-Vandenbergen, 2003; Fox Tree, 2010; Roze et al., 2012; Heine, 2013; Beeching, 2019; Rysová, 2017; Crible & Degand, 2019). However, they remain mysterious and pose a real challenge to be taught to foreign language learners.

Studying discourse markers in oral conversation is central to understanding how discourse markers behave. However, oral corpora are difficult to gather or create and always present some biases. Different possibilities exist, and some studies chose to focus on subtitles, as they are an easily accessible source for parallel corpus. Earlier, research was mainly focused on film subtitles (Chaume, 2004; Biagini, 2010; Furkó, 2014; Degand, 2015; Levshina, 2016) but more recently, as the cultural offer broadens, some authors have chosen reality TV shows as a base for their studies (Zenner & Geeraerts, 2015; Zenner & Van De Mieroop, 2017; Ro & Jung, 2022). However, to the best of our knowledge, there are no studies comparing the two media.

This study proposes a contrastive approach to investigate discourse markers in subtitles corpora. Three discourse markers have been selected: ‘you know’, ‘well’ and ‘so’, as being the most frequent in the corpus of subtitles. Two main criteria have been selected to conduct the analysis: the annotation system developed by Crible and Degand (2016) and the position of the discourse marker in the sentence. The first part of the study focuses on comparing the usage of the three discourse markers in film and reality TV dialogues. These two formats present similarities but differ greatly in some aspects, such as preparation or spontaneity. The aim is always to be as close as possible to natural speech. The second part of this study aims to analyse the different translation choices in French and Dutch done by the translator, and when he chooses not to translate the discourse marker. From previous studies (Degand, 2015; Connors, 2016), more than 70% of the discourse markers can be lost in translation. However, it

seems that it also depends on the discourse marker and the target language (Furkó, 2014; Crible et al., 2019). The objective is also to see if a pattern can emerge regarding the discourse marker or the coherence relation it conveys. Abuczki et al. (2018) identify the sequential domain as the least likely to be translated.

The first part of this study focuses on the state of the art to acknowledge what has already been done in the field and the results demonstrated by other authors. The second part introduces the different corpora used for this study and the procedure adopted for assembling them. The methods part also introduces different criteria chosen for this analysis: the annotation system (Crible & Degand, 2019) and the position of the DM in the sentence. The following part exposes the results, comparing the usage of the three selected discourse markers in film and reality TV dialogues and the contrastive analysis of the translations in French and Dutch. Finally, the limitations and further studies are debated in the discussion and the study ends with a conclusion.

2 State of the art

2.1 Discourse markers

Discourse markers (henceforth DMs) have been a subject of research for many years. They are also known as discourse connectives (Rouchota, 1996; Roze et al., 2012; Rysová, 2017), pragmatic devices (Stubbe & Holmes, 1995) or pragmatic markers (Fraser, 1996). Their denomination however has changed with the year and/or the author. They are present in every language (Biagini, 2010) and occur beyond the propositional content of the communication (Fraser, 1996; Fox Tree, 2010).

DMs are difficult to define, as they have many roles. Schiffrin (1987, p. 31) described them as "sequentially dependent elements which bracket units of talk". A general definition which allows the two different categories, the 'connectives' (*so, but, because…*) and the 'pragmatic particles' (*well, you know…*) to be integrated (Crible & Degand, 2019). Schiffrin's definition is not very specific and allows for multiple interpretations. For instance, whether interjections have to be considered as DMs or not, varies from author to author (Heine, 2013). The debates about the categorial definition of discourse markers are still ongoing.

Fox Tree (2010, p. 3) defines discourse markers as "conventionalised, learned expressions that provide information about how the propositional content of messages should be interpreted".

According to Crible and Degand (2019, para. 1) "in human communication, discourse is where the magic happens. It is through markers of structure and interaction that speakers convey not only the coherence of their intended message but also their attitude towards this message and towards the interlocutor."

DMs present some common characteristics. They are syntactically flexible and non-obligatory (Bosker et al., 2021). They also play a crucial role in unprepared discussions (Fox Tree, 2010). They facilitate hesitations, reformulations, seeking agreement with the interlocutor and more. These elements occur in everyday conversation and make the speech more natural. Nearly all may occur in sentence-initial position, cannot be negated and they normally can be omitted without losing grammaticality or propositional content (Fraser, 1996).

The roles of discourse markers are varied and numerous. They are used to convey an utterance of a speaker-hearer interaction, speaker attitudes and/or the organisation of texts (Heine et al., 2021; Takamura, 2020). They take part in assisting in turn-taking, contributing to social solidarity. Their roles can even be contradictory as turn-initiator and turn-relinquisher (Fox Tree, 2010).

DMs are also an intriguing subject for L2 (second language) learners. They are challenging to acquire in a classroom. Teaching DMs is particularly difficult as their usage is unique and situational. The usage of discourse markers is specific to a language, and even if there are some literal equivalents, they are not always used in the same way. The acquisition of DMs is mainly pursued through imitation, thanks to interactions and communications with native speakers. There is consequently a direct link between the acquisition and native-like use of discourse markers and language proficiency (Tsai & Chu, 2017). Moreover, discourse markers do not bear any strong meaning content and can be removed without altering the global understanding.

For example, (from *Once upon a time in Mexico*)

(1)

"(…) after he's killed the president. Savvy?	"(…) après qu'il aura tué le président. Pigé ?	"(…) hij de president heeft vermoord. Snap je?
- **So** why me?	- Pourquoi moi ?	- Waarom ik?
- You've got nothing to live for ... and, in a way, you're already dead."	- Vous n'avez aucun but dans la vie ... et, quelque part, vous êtes déjà mort."	- Jij hebt niets om voor te leven ... en je bent eigenlijk al dood."

In this example, there is a 'so' in the English original line and the translator chose not to translate it into French or Dutch. We can see here that the meaning remains similar in the three languages.

Their value is crucial to express subjectivity or interactivity during exchanges (Biagini, 2010).

(2)

"**Well, well**. What does it say?"	"**Alors**, ça dit quoi ?"	"**Nou**, wat staat er?"

In this example from the film *Gone Girl*, the ‘well’ is repeated and used for interactivity. We can see a translation both in French and Dutch to maintain this emphasis.

2.1.1 WELL

According to Aijmer and Simon-Vandenbergen (2003) and Cuenca (2008), ‘well’ is one of the most analysed discourse markers. Schiffrin (1987) studied ‘well’ as a marker of response. She identified the roles of ‘well’ as used to fill, hesitate, change topics or as a turn-initiator. This view of ‘well’ as a marker of response has led to consider it as a politeness marker for signalling an unexpected or unfavourable answer (Takamura, 2020).

‘Well’ has been studied in various contexts, in New Zealand courtrooms (Innes, 2010), in films with Catalan and Spanish subtitles (Cuenca, 2008) or Vietnamese subtitles (Diem, 2023), from multilingual corpora with Dutch and Swedish equivalents (Aijmer & Simon-Vandenbergen, 2003), and in Chinese-speaking learners corpora (Huang, 2018), to cite a few.

‘Well’ can convey different pragmatic meanings and is therefore difficult to be determined (Cuenca, 2008). It appears important to differentiate the use of ‘well’ as an adjective or adverb and as a discourse marker (Takamura, 2020). ‘Well’ is an adverb in the following example from the film *Bridge of Spies*:

(4) I'm sure that you would wish them to be treated **well**.

‘Well’ is used as a derivation of the adjective ‘good’ to add a characteristic to the verb ‘treat’. To contrast, ‘well’ in the initial position is most of the time a discourse marker, as an illustration in (5) from the same film:

(5) **Well**, I said no.

The use of ‘well’ does not add any semantic content to the sentence, and therefore it can be removed without altering the general meaning.

2.1.2 SO

Schiffrin (1987) defines ‘so’ as a marker of cause. She argues that ‘so’ implies an inference between the propositions, making the proposition introduced by ‘so’ as the main idea unit. Raymond (2004) in particular, studied the ‘stand-alone so’. In his

analysis, he identified that the ‘stand-alone so’ implies a consequence link in an unfinished turn. It hints the interlocutor to reconstruct a logical link to what has already been said.

Buysse (2012) investigates the different functions of ‘so’ and identifies 10 of them. ‘So’ is a particularly versatile discourse marker. From his study, it appears that ‘result’ is the most common use of ‘so’, but it can also be used as conclusion, summary, or prompt amongst other functions. In another study (2014) Buysse focuses on final ‘so’ used as ‘non-prefatory’. This final ‘non-prefatory so’ has a “transition-relevance place, i.e., a point at which the turn may be shifted to an interlocutor” (Buysse, 2014, p. 31). It is left to the interlocutor to draw an inference or to take their turn, or both.

An example of turn-yielding role of the ‘stand-alone, non-prefatory so’ from *Love is Blind*:

(6) My dog sleeps with me every night, **so**...

The consequential link must be reconstructed, hinting to the interlocutor that they can take their turn.

‘So’ is also used as a topic changer (Bolden, 2009). An illustrative example from *Love is Blind* of ‘so’ being a topic changer:

(7) -It feels so good to hear your voice again.

-Yeah.

-**So**, something, seemed like in your past, or something has happened

that has caused you to feel that you're not beautiful.

2.1.3 YOU KNOW

In her study, Schiffrin (1987) identifies the role of ‘y’know’ as participation and information. According to her, ‘y’know’, more than any other discourse markers, is greatly influenced by its original semantic meaning, i.e., you and know, hinting directly to the information field. Someone (the ‘you’) has information about something (the verb ‘know’). Its interactional role is also clear, as the ‘you’ refers to the interlocutor.

Stubbe and Holmes (1995) studied ‘you know’ in New Zealand English. They focused mainly on the social aspect of this discourse marker, underlining the ability of ‘you know’ which tends to emphasise solidarity and implicit shared understanding. ‘You know’ is also partly negatively perceived and socially sanctioned.

Fox Tree and Schrock (2002) investigate the roles of ‘you know’ and ‘I mean’ in detail. According to them, ‘you know’ serves as positive politeness and a repair marker. It lowers the social distance and allows a more casual speech through a desire for shared experience. ‘You know’ is also used to control the fluency of speech and to stall for time. It has moreover a role of turn management or monitoring. The last role is organisation as an enquoting device or as a topic changer. They conclude by stating that ‘you know’, by its nature, encourages the addressee to dive into their own thoughts.

According to Beeching (2019), ‘you know’ creates a common ground. The position of the discourse marker may also influence its function. In first position, the function of ‘you know’ is attention-getting oriented, while in middle position it is mainly used as a pause filler. When used in last position ‘you know’ often fulfills the role of agreement seeker.

Buysse (2017) studies ‘you know’ in EFL speakers and establishes that ‘you know’ also verifies if the interlocutor shares the knowledge needed to understand the prior proposition. ‘You know’ works as a politeness marker, preventing both parties from losing face, seeking common agreement instead of tagging one part as ignorant. It can be also used as an empathetic device, when not hinting at a certain knowledge but more to a situation the interlocutor can relate to.

2.2 Discourse markers in subtitles

The use of subtitles as a valuable resource for linguistic research is an interesting aspect of the present study. Biagini (2010) examined discourse markers in French films and their translations in Italian subtitles. She states that subtitles are a good alternative to spoken corpora because subtitles are meant to be written to give a feeling of authenticity. They are, moreover, easily accessible. Biagini chose to study discourse markers as their value is crucial for expressing subjectivity or interactivity during communication. She concludes that subtitles are a good source of "written verbalised

speech" ("discours oralisé écrit") (Biagini, 2010, p.31) and, therefore, good material for studying discourse markers as they appear frequently in spoken language. However, her finding reveals that many discourse markers were not translated, leaving the role to the image and/or context to convey the missing information. This could be attributed to the restrictions of the written form of subtitles, which must conform or adhere to specific rules.

Karankata et al. (2020) recently investigated these restrictions in that they cannot be too long (i.e., take more than 10% of the screen) and are small enough for the audience to read before the scene ends. Furthermore, they should maintain a linguistic whole and be aesthetic. Aestheticism is a subjective criterion, but it concerns the length and position of subtitles to enable the audience to read them easily.

In her study, Degand (2015) demonstrates that a large majority of discourse markers are lost in translation. She investigates the proximity of subtitles to natural speech. To prove that point, she looks at the discourse markers, as the more numerous these are, the more natural the speech. She also shows the extent to which discourse markers remain in the translated subtitles. Her results reveal that only 37% of the discourse markers were translated. However, the translator also adds discourse markers in the subtitles, which were not present in the original English conversation. Based on her analysis, it was found that such additional discourse markers comprise almost 50% of the total discourse markers in the translated version of the subtitles.

Connors (2016) examines ‘enfin’ and ‘écoute’ and their translations to English in French films and series. She uses nine French films and the four first episodes of two French series for her corpus. Out of the 119 occurrences of ‘enfin’ and 48 occurrences of ‘écoute’, she accounts for 72% of omissions for ‘enfin’ and 73% for ‘écoute’.

In his study, Furkó (2014) investigates the DMs ‘actually’ and ‘I mean’ and their Hungarian counterparts. He shows that out of 288 tokens of ‘actually’, 20% were lost in translation. Of 133 tokens for ‘I mean’, 26% were not translated. Even if the sample is limited, he also analyses a Hungarian counterpart, ‘vagyis’, and it appears that 43% of the occurrences of ‘vagyis’ are added by the translator where there was no discourse marker in the original English text.

Crible et al. (2019) analyse different discourse markers in English and their translations in Lithuanian, Czech, French and Hungarian. Their results show that the chance to be

translated varied from language to language. For example, 30% of the DMs were not translated in Lithuanian, 38% in Czech, 44% in French and more than 50% in Hungarian (Crible et al., 2019, p. 144).

Abuczki et al. (2018) focus their study on the DM ‘and’ in TED talks. They use English and five languages subtitles, Lithuanian, Czech, French and Hungarian. Out of the four pragmatic domains, they show that the sequential domain is the more likely to be omitted.

Levshina (2016) works on verbs of letting in Germanic and Romance languages. In her study, she identifies a phenomenon typical to translations: "translationese". Translationese could be a bias due to the influence of the original language. According to her article, if a structure is present in the original language, the number of literal translations of this structure tends to be higher than for a native speaker.

In his study, Chaume (2004) investigates the impact of the loss of discourse markers in subtitles. His results demonstrate that the texts in subtitles are less cohesive due to the suppression of discourse markers. The understanding, however, is not impaired because the spectator automatically assumes that there are semantic relations between the sentences. The audio-visual context also provides some hints to help to grasp the underlying meanings. According to Chaume (2004), the translation in subtitles forms a genre by itself, and the addressees recognise its characteristics and assume that it is less coherent than the original lines. At last, the audio-visual text is a redundant text, which allows the spectator to easily understand a less coherent discussion in their native language. The semantic meaning remains the same in the subtitles, but the interpersonal meaning is hindered.

2.3 Captions or subtitles?

Several papers insist on the difference between captions and subtitles. The field of these articles is mainly the didactics of English as a foreign language (EFL), for instance, Hosogoshi (2016), Pujadas and Nuñoz (2020), Vanderplank (2013) or Peters et al. (2016), to cite a few. Captions provide the viewer with the script of the dialogues and are primarily used for the deaf and hard-of-hearing (Peters et al., 2016). Subtitles are on-screen text in the viewers’ native language (Hosogoshi, 2016). This distinction, however crucial for the EFL studies, bears less interest in the frame of the present

study. The captions in English are here always the starting point, whereas the subtitles in French and Dutch form the second part of the analysis.

2.4 Reality TV show VS films dialogues

The first part of this study focuses on identifying potential differences between film and reality TV show dialogues. As stated before, film dialogues are written to mimic real conversations. Reality TV show, however, supposedly lacks this scenario. In their study, Zenner and Geeraerts (2015) state that the reality TV show they investigate (*Expeditie Robinson*) forms “a corpus of naturally occurring spoken conversation” (Zenner & Geeraerts, 2015, p. 247). To assess that reality TV shows are not scenarised or prepared at all is maybe taking a step too far, but the lack of prepared written speech is surely an element worth considering. Ro and Jung (2022) speak of “relatively unscripted TV programmes” (Ro & Jung, 2022, p. 89) in their study about a Korean reality TV show (*Babel 250*).

For the present study, the reality TV show *Love is Blind* has been selected. In this ‘social experiment’, fifteen men and fifteen women date each other in “pods”. These pods allow them to have a conversation without seeing each other. After ten days of dating, they may propose to each other and then finally see the face of their promised one. The show follows the new couples on a pre-marriage holiday and in some reconstructed day-to-day life in a flat. They meet friends and family before going to the altar. They have then a last chance to cancel the wedding by saying ‘no’.

The reality TV shows have several advantages apart from the participants having different social backgrounds. In comparison with other reality TV shows such as *The Bachelor* or *Too hot to handle*, the participants are here to find a life partner and not to become famous. The absence of votes or money prize help to prevent exaggerated acting and well-rehearsed speeches. In the same way as films, the subtitles for reality TV shows are easily accessible on the web and can provide a consistent source for corpora.

The dialogues consist of different parts. The female and male participants are initially placed separately in two separate common rooms. The male participants and the female participants are allowed to interact through a process of ‘dating in pods’. Here they ask personal questions, get to know each other, and discuss each other’s interests

without seeing each other's faces. Then, they have a few face-to-camera moments where they recall and comment on their past conversations in pods. Between the dates, they also converse and discuss, with the other same sex participants in their respective common rooms. Later in the show, they also have discussions with friends and families.

The downsides of reality TV shows are that they have been edited, cut, and pasted to form an exciting episode with moments of deep feelings. The logic and the sequence of the dialogues may also have been altered by editing and cutting (Zenner & Van De Mieroop, 2017). This bias is particularly important to be accounted for in this study, as the analysis of discourse markers is directly linked with the different sequences of speech. There is then no guarantee that the final broadcasted, reconstructed sequences are in chronological order.

2.5 Research questions and hypotheses

Based on the above literature survey, it is necessary to answer the following three primary research questions:

1. How does the use of discourse markers differ between film dialogues and reality TV shows?
2. To what extent are discourse markers translated in subtitles?
3. What factors influence the translation choices of DMs in subtitles?

Further we will be addressing the following secondary question:

- Can a contrastive analysis of the translations of certain DMs reveal common patterns among different languages and shed light on the translator's choices?

We can draw some hypotheses from these questions:

1) The dialogues from the films are supposed to be as close as possible to real-life conversations (Biagini, 2010). However, these dialogues are still written, prepared, and rehearsed. In contrast, reality TV show conversations are unscripted (Ro & Jung, 2022). They are nevertheless not without biases, as Zenner & Van De Mieroop (2017) stated. The conversations are edited, cut, and adapted, and the final edition is not guaranteed to be chronologically

reconstructed. A difference between the two is than expected, but no comparative studies have been conducted yet.

2) A gap is obvious between the subtitles and real conversations, mainly due to the rules and the particular environment of subtitles (Karankata et al., 2020). According to Degand (2015), it is expected that around 60% of the discourse markers are not translated. Following Connors's results (2016), around 70% of the discourse markers were not translated. A large majority of DMs are then expected to be lost in translation.
3) Supposing that between 60% and 70% of the discourse markers are not translated, the remaining DMs must be important compared to those eliminated. Chaume (2004) demonstrates that this loss of discourse markers does not hinder the understanding and the coherence relations are left to the audio-visual scene or to the spectator to be reconstructed. Based on that, some relations may be easier to implicitly convey than others. Abuczki et al. (2018) study the DM 'and' and account that the sequential domain is less likely to be translated.

3 Methods

3.1 Corpus-based approach

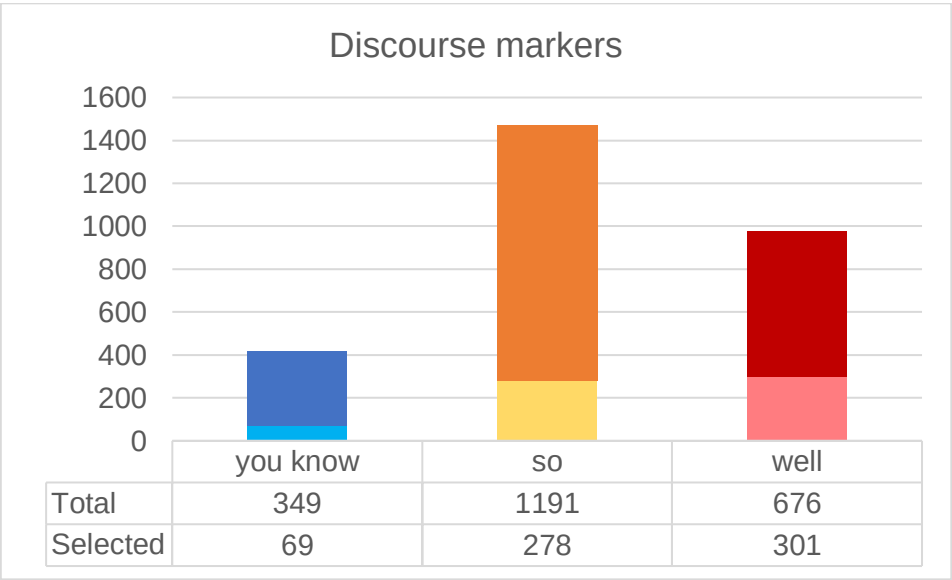
This study will mainly adopt corpus-based approaches to answer the above-stated research questions. For this purpose, the following two parallel subtitle corpora have been used to create a newly assembled corpus:

1. ParTy (Levshina, 2016) – 11 films
2. OPEN Subtitles (Tiedemans, 2012) – 26 films

These two corpora contain captions from English films and their subtitles in French and Dutch. This choice was mainly motivated by the fact that the second part of the study is a contrastive analysis, and these corpora are easily accessible. They are selected after ensuring sufficient data is available for analysis in these three languages.

Through database analysis, it has been found that among other common DMs, 'you know', 'so', and 'well' are the most frequent. Therefore, only these three DMs have been considered for this entire analysis. The final data selected for the analysis are manually confirmed to satisfy the following three criteria:

(i) They are definitely a discourse marker
(ii) Equivalent translations are available in both French and Dutch
(iii) Translations are semantically not too far from the original sentences



	you know	so	well
Total	349	1191	676
Selected	69	278	301

Figure 1. Distribution of 'you know', 'so' and 'well' in the assembled corpus.

Based on the above criteria, only 19% of ‘you know’, 23,3% of ‘so’ and 44,5% of ‘well’ has been selected for the analysis. The detailed distribution of these DMs in the assembled corpus is presented in the form of a histogram in Figure 1.

The following table provides information about the database considered for this study.

Table 1. Total number of films and the distribution of DMs in the assembled corpus.

S.No	DMs	Total number of films	Total number of discourse markers
1	You know	37	69
2	So	37	278
3	Well	37	301

3. *Love is Blind* (LiB) (Netflix, 2020)

A new corpus has been compiled by extracting subtitles from the popular Netflix reality TV show *Love is Blind*. The subtitles from the first seven episodes of season one have been utilised for the analysis. This American reality TV show stages single men and women who enrol themselves into a ‘social experiment’. The ultimate objective is to find the love of their life. They initially converse, virtually go on dates, and if everything goes well, they eventually get engaged before meeting in person. The show continues after the couples meet and follows them until the wedding ceremony, where the contestants are given one last opportunity to refuse the union. This show has been selected due to the nature of the participants, excluding experienced ‘stars’ from other reality tv shows or professional actors in the making, and the absence of a scenario and pre-prepared speech or written dialogues.

For the sake of comparison, the same discourse markers, i.e., ‘well’, ‘you know’ and ‘so’ considered in the assembled corpus, have been selected. The same criteria have been adopted, but only the first criterion was relevant for this new corpus (i.e., the phrase considered is actually a discourse marker). The subtitles extracted from this reality show were always complete without any missing lines, which was often the case in the assembled corpus. Further, the translations here were never too far from the

original. This is probably because the context is relatively simple due to the informal register of the conversations, which is often the case in any reality TV show.

This new corpus contains 57,088 tokens from which all occurrences of the three DMs mentioned earlier have been extracted. The details are presented in the following Table 2.

Table 2. The total number of episodes and the distribution of DMs in the LiB corpus.

S.No	DMs	Total number of episodes	Total number of occurrences
1	You know	7	167
2	So	7	243
3	Well	7	60

3.2 Procedure

The three corpora, ParTy, Open Subtitles, and the new 'LiB', have been extracted using the three procedures described below.

3.2.1 Corpus ParTy

In order to extract the data from the subtitle corpus ParTy, the following procedure has been adopted:

- The subtitle corpus can be accessed through GitHub (https://github.com/levshina/ParTy-1.0).
- For every film (originally in English) in the database, separate text files with the corresponding subtitles in user-specific language are available for search and download.
- The text files containing the line number, the English text and its subtitle translation in French and Dutch can be accessed separately.
- Keeping the line numbers as the reference, a manual search has been performed in GitHub for the specific DMs of our interest.
- The relevant phrases filtered through the search are extracted one by one manually after confirming that they satisfy the aforementioned criteria. A similar procedure is followed separately for the text files containing English to French

translations and English to Dutch translations. Finally, a database has been created using the commercially available software: Microsoft Excel.

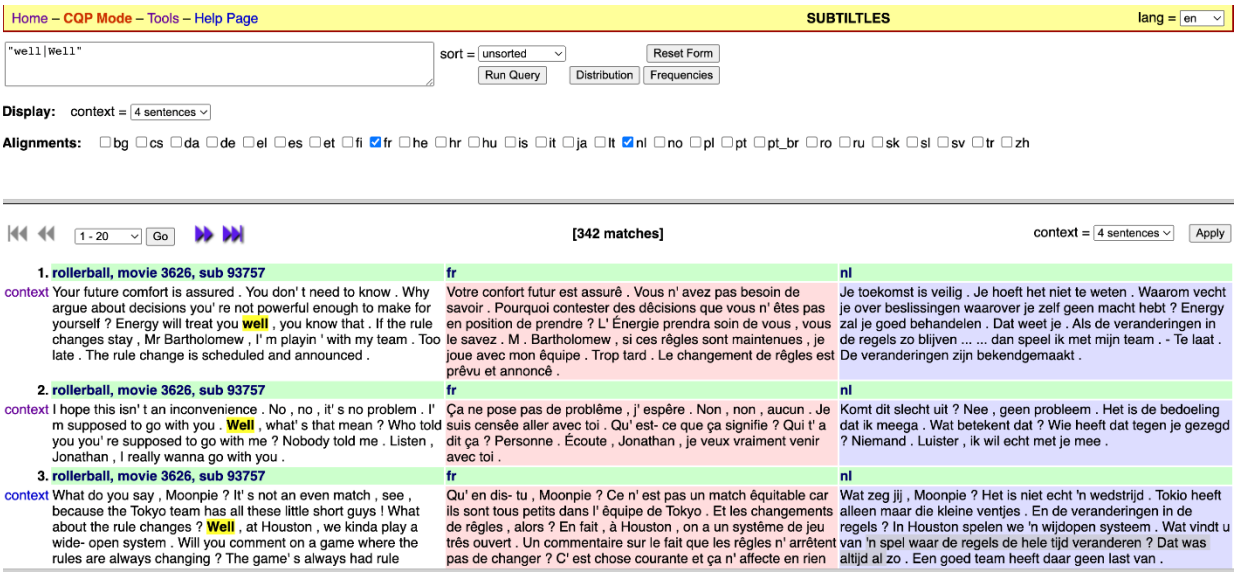
- The analysis, such as word counts, frequencies, and other valuable classifications (explained later in detail), have been manually coded.

3.2.2 Open Subtitles

In order to extract the data from the Open Subtitles corpus, the following procedure has been adopted.

- The subtitle corpus was accessed using the subtitle search interface provided on the main OPUS website.
- Unlike ParTy, there is an inbuilt search engine that enables the user to choose the original subtitle language (lang=en), the word that needs to be searched (‘you know’, ‘well’ and ‘so’), display context (i.e., number of words, number of sentences and so on) and Alignments (in the present case it is French (fr) and Dutch (nl)).
- Note that the search words are case-sensitive. So, it was necessary to specify all the possibilities of occurrence separated by logical OR symbol, as shown in Table 3.
- If a query is initiated with all the necessary options specified, a well-aligned database with three columns, English, French and Dutch, appears with the search word highlighted in English.
- Another advantage of this search engine is that helpful information, such as the frequency and distribution of the search word, can be directly obtained.
- Similar to the procedure described for the ParTy corpus, all the extracted data is organised and assembled into a single excel file for further analysis.
- This database will be hereafter referred to as the ‘Assembled Corpus.’

Table 3. OPUS search engine interface.



3.2.3 LiB corpus

In order to extract the data from the subtitle corpus LiB, the following procedure has been adopted.

- The subtitles from the first seven episodes of the first season of *Love is Blind* have been accessed using a personal Netflix account. A simple procedure described below is used to extract all the necessary data.
- Netflix has been accessed with active credentials using google chrome as a web browser.
- The developer tool window is opened using the shortcut key: CTRL + Shift + I. Once the appropriate subtitles are loaded, the subtitle data will appear in this network tab. The subtitles data can be accessed through a simple search with the token '?o='.
- Once the relevant episode is played and paused on Netflix with the appropriate subtitles, the necessary subtitle data, along with their timestamp, can be downloaded as Extensible Markup Language (XML) file through the network tab.
- The English captions are extracted first, followed by French and Dutch subtitles. Finally, the XML files are converted to Comma separated Variables (CSV) files using an online tool.

- From the CSV files, all occurrences of ‘well’, ‘so’ and ‘you know’ were manually selected and along with the files containing the corresponding translations in French and Dutch. All these data have been ensembled into an independent Excel file for further analysis.
- This database will be referred to as the ‘LiB’ corpus.

3.3 A two-dimensional account of discourse markers

For the present study, the annotation system developed by Crible and Degand (2019) has been adopted, focusing on discourse markers instead of relations. This system identifies four domains and fifteen functions as explained in Section 4.1. This approach is based on the Penn Discourse Treebank (henceforth PDTB). The PDTB is one of the most widespread discourse annotation protocols. The latest version is the third (3.0), as detailed in Prasad et al. (2018). This hierarchical taxonomy organises the different discourse relations into categories with different levels (see Figure 2). It investigates the relations between the first (Arg1) and the second argument (Arg2) as the base for the taxonomy.

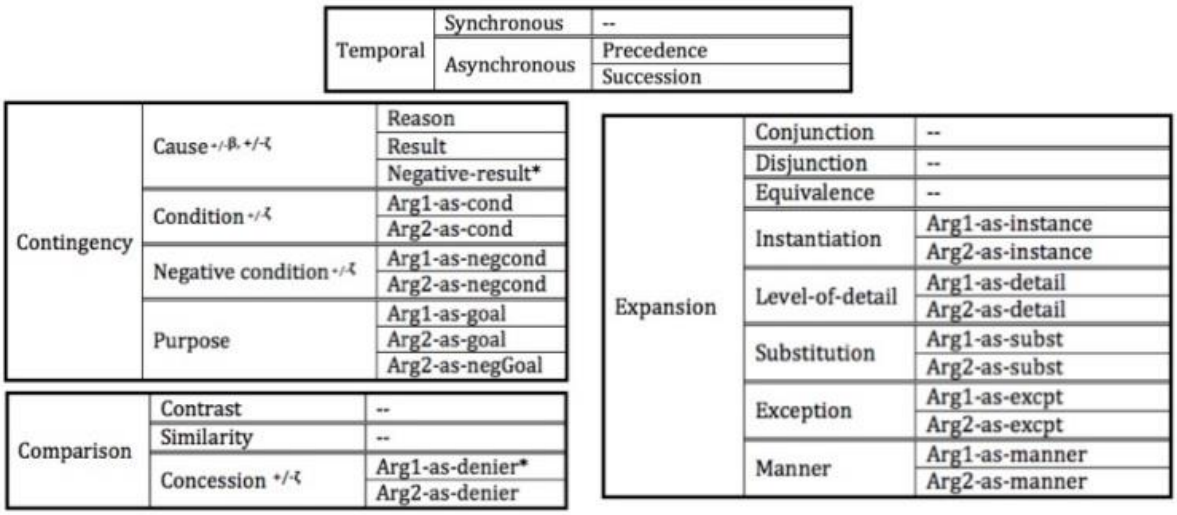


Figure 2. PDTB-3 Sense Hierarchy (Prasad et al., 2018, p. 90).

Crible and Degand (2019) reworked the semantic functions and added the four domains: viz., ideational, rhetorical, sequential, and interpersonal, for each possible function.

4 Results

4.1 Semantic functions and pragmatic domains

The first part of this analysis focuses on comparing the data from film dialogues and reality TV show. For the analysis, two main criteria have been selected: the semantic and pragmatic values of the DMs and their position in the sentence.

The values of the discourse markers are analysed using a unique annotation system developed by Crible and Degand (2019). This system utilises the dual approach for both semantic and pragmatic annotations. The system defines 15 different semantic values, namely, addition (ADD), alternative (ALT), cause (CAU), concession (CCS), condition (CND), consequence (CSQ), contrast (CTR), hedging (HDG), monitoring (MNT), specification (SPE), temporal (TMP), topic (TOP), quoting (QUO), agreeing (AGR) and disagreeing (DIS), and four pragmatic values, viz., ideational (IDE), rhetorical (RHE), sequential (SEQ) and interpersonal (INT). The analysis of the position is carried out with three possibilities: first, middle or last position. The aim is to assess if any significant difference can be identified between the written dialogues mimicking natural speech and the conversations recorded in a reality TV show.

The coding has been conducted manually on Excel. For each discourse marker one semantic value out of the fifteen has been given. A decision was made regarding the context, and in case of ambiguity, the original source was then exploited. Then a pragmatic domain was added following the same method. The process has been done twice, hiding the first coding while conducting the second one. The interrater agreement for the semantic values is 0,77 and for the pragmatic ones 0,74. For the final decision, both the original source (the film or the reality TV show) and the two translations were used to assess which choice was the more accurate. The difficulty of the coding resides in understanding the intention of the speaker. As stated before, the dialogues in reality TV shows are reconstructed and edited, so the context is probably sometimes biased and then difficult to decipher. For the films, the range is more extensive, as the writer had time to think and consider which DM to use in which context, but it is still clearer. The interrater agreement was consequently higher for the films than for the reality TV show with 0,79 for the films for both values and for the reality TV show, 0,75 for the semantic functions and 0,71 for the pragmatic domains.

As for the position, the same process has been adopted: a manual double coding. The agreement rate is higher (0,94) as hesitations were less frequent. Unsurprisingly the interrater agreement is higher for the films (0,97) than for the reality TV show (0,90). The difficulty lies in the reconstructed dialogues, as a decision was to be made if the discourse marker had a link with the previous sentence or with the following, which was not so obvious all the time, especially in the reality TV show.

The distribution of semantic values in film dialogues and in the reality TV show are presented in Figures 3 and 4, respectively. For the film dialogues, the semantic values are mainly topic (TOP), monitoring (MON), specification (SPE) and consequence (CSQ) (see Figure 2).

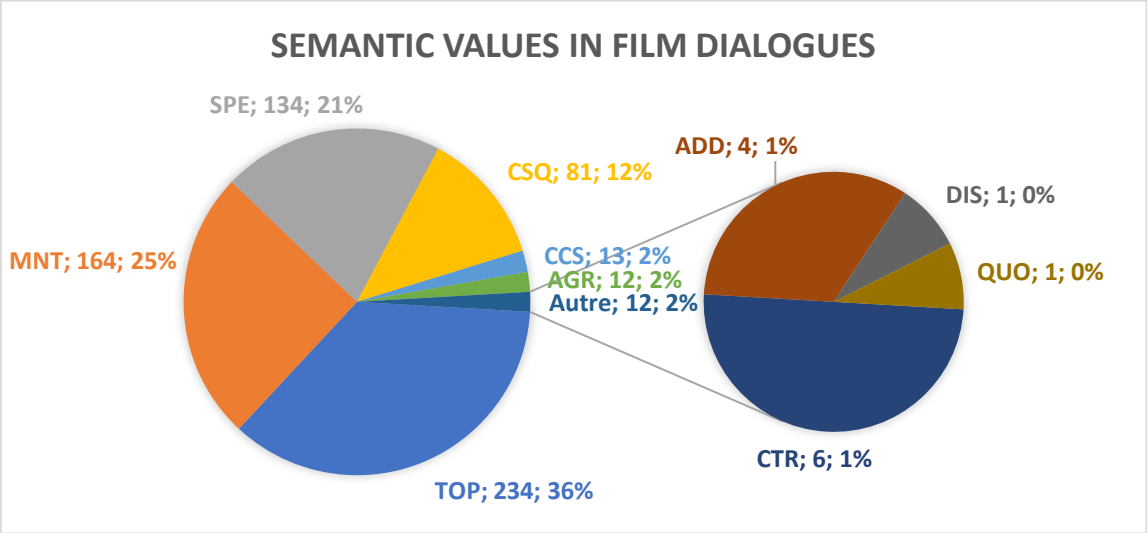


Figure 3. Distribution of semantic values in film dialogues for the three DMs.

In the reality TV show, the main semantic values are monitoring (MNT), consequence (CSQ), topic (TOP) and hedging (HDG), as shown in Figure 4. Combining the most common semantic values of films and reality TV show, we end up with five semantic functions: topic, monitoring, consequence, specification, and hedging. These values will be mainly discussed in the rest of this study. Since the occurrence of the other values is insignificant, the remaining values will be treated as unrepresented or exceptional. Further, generalisation for such values is also difficult to establish with such few occurrences.

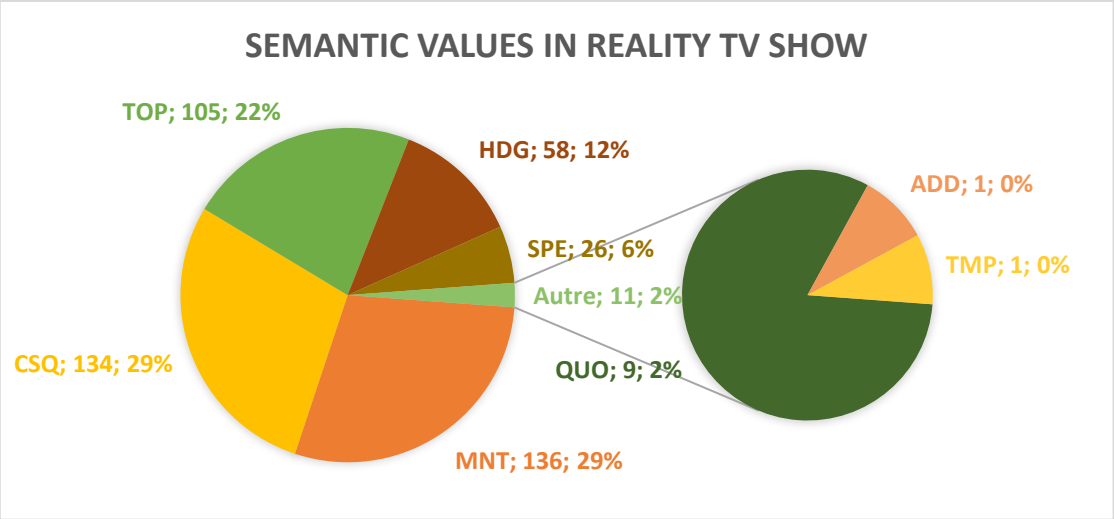


Figure 4. Distribution of semantic values in reality TV show for the three DMs.

The distribution of pragmatic values for film dialogues and reality TV show is presented in Figure 5. The proportion of the pragmatic values follows a similar scheme. Sequential (SEQ) is the main value, while ideational (IDE) is the least represented. However, the proportion of ideational and rhetorical (RHE) is higher in film dialogues than in reality TV show, where the proportion of interpersonal values is more important.

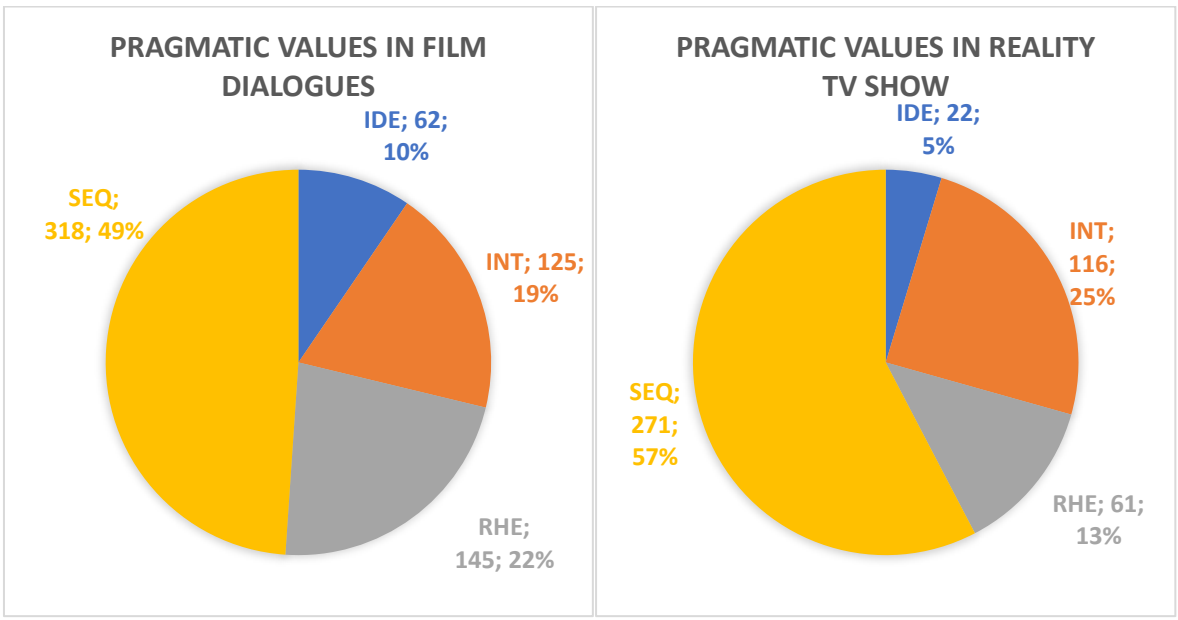


Figure 5. Pragmatic values distribution in film dialogues and reality TV show for the three DMs.

The three discourse markers selected for this study will be analysed case by case in the following sections to specify the differences between film dialogues and the reality TV show based on the occurrence position and the values.

4.1.1 YOU KNOW

The results for ‘you know’ show that the distribution occurs in equal proportion for the film dialogues. This could be explained by the fact that the dialogues are written, revised and rehearsed several times. This allows more control over the position than an unprepared, real-time conversation. As seen in Figure 6, the scriptwriters seem to balance the positions of ‘you know’. It is, however, different for the reality TV show. The middle position is dominant, where it enables the use of filler to stall or hesitate. The use of ‘you know’ at the end of a sentence to keep the conversation flowing seems slightly underused compared to the film dialogues. An example of ‘you know’ at the end of the sentence can be observed from the film *Scarface*:

(8) You tell me, Tony. Frank is smart, **you know**? You can't blame him for that animal. It's a crazy business we're in, **you know**? That could happen to anyone, even you. Why don't I go back and talk to Frank ... and work it out? I fix things between us.

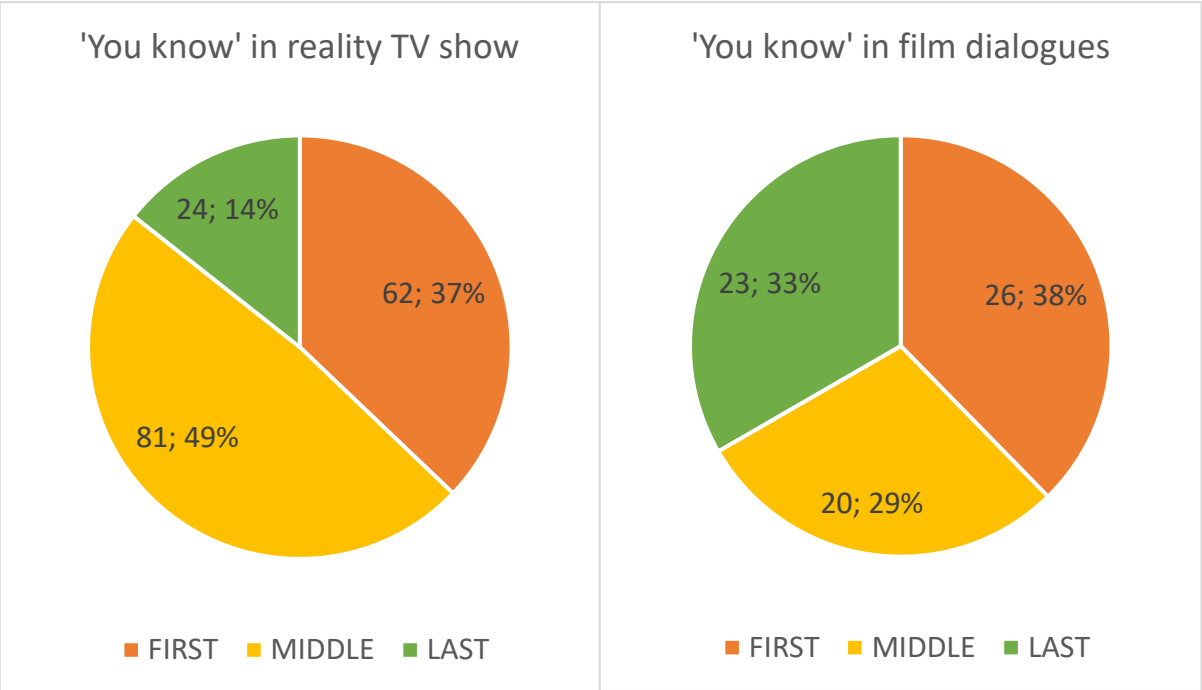


Figure 6. Position of 'you know' in reality TV show and film dialogues.

The primary use of ‘you know’ is monitoring, as shown in Figure 7. The number of occurrences is higher in the reality TV show. A correlation can be made between the main middle position of ‘you know’ and the sequential monitoring (MNT SEQ) when the DM is used as a mark of punctuation to hesitate or stall for time (Crible & Degand, 2019). This observation can be made from an example from the LiB corpus:

(9) Today, I feel strongly towards two people, but I, **you know**... Like, I try to avoid that conflict as well, I guess.

It seems that ‘you know’ is underrepresented as a hesitation marker in written dialogues compared to a free or less formal context of the reality TV show.

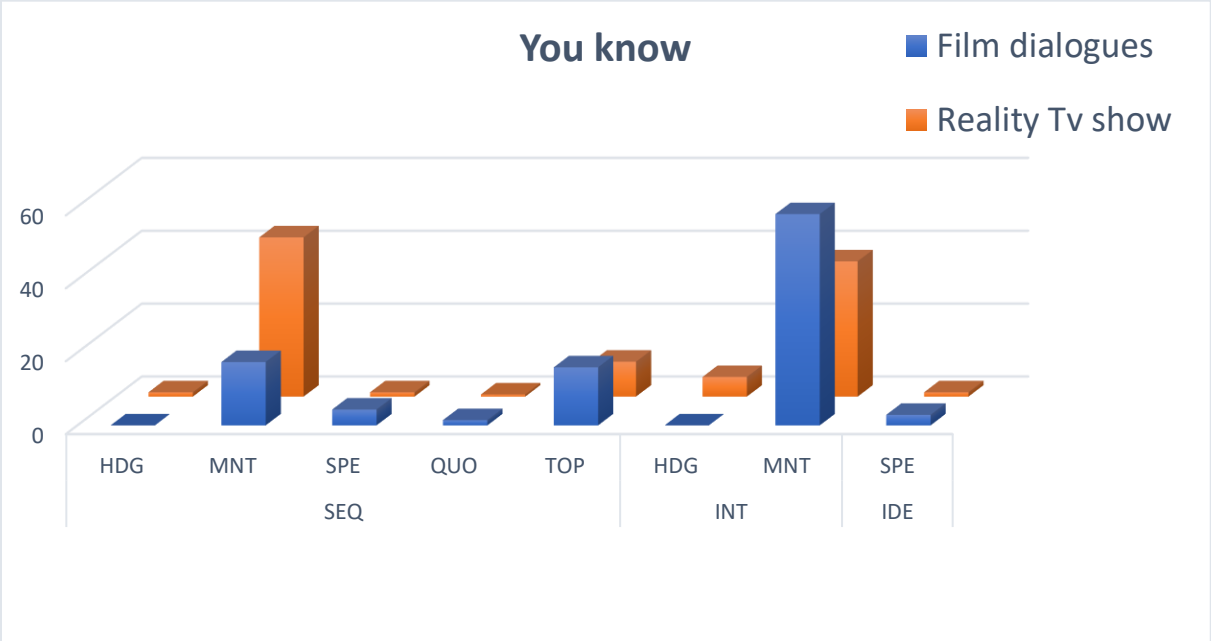


Figure 7. Semantic and pragmatic values of 'you know' in film dialogues and reality TV show.

4.1.2 SO

As illustrated in Figure 8, the position of ‘so’ seems to follow a similar pattern in the reality TV show and film dialogues. In both cases, ‘so’ is highly likely to occur at the beginning of the sentence.

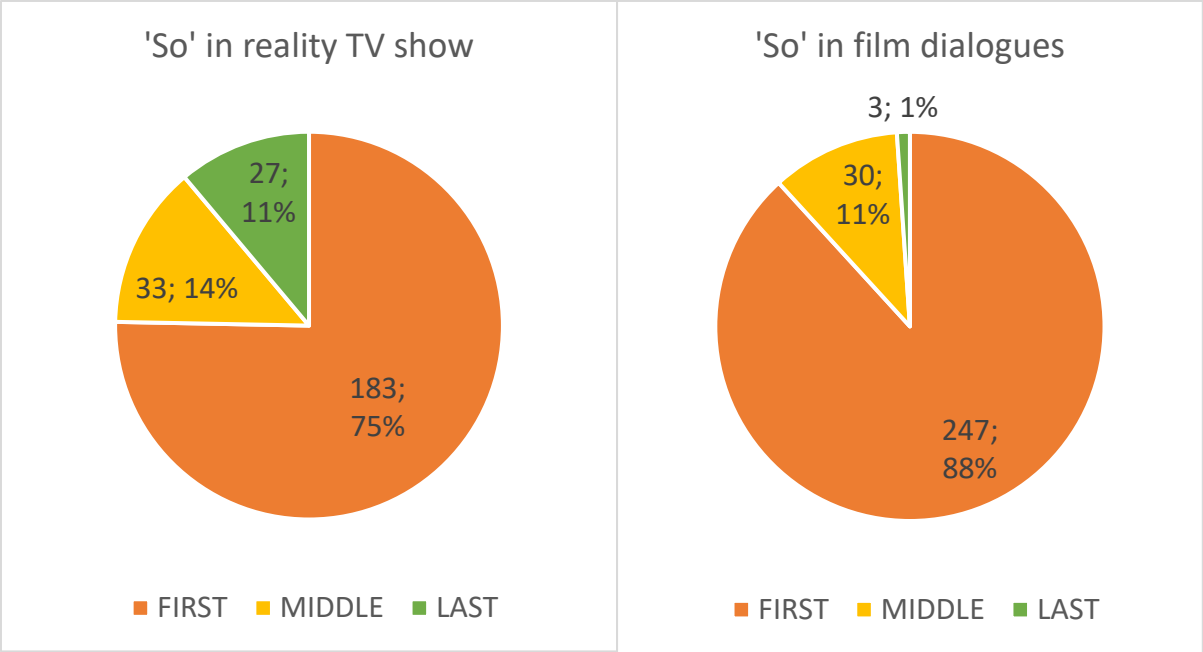


Figure 8. Position of 'so' in reality TV show and film dialogues.

'So' is often used as a consequence marker (CSQ), even if it can occur at the beginning of the sentence, without any semantic value other than changing the topic (TOP), as seen in Figure 9. The relative absence of 'so' as the last item of a sentence in film dialogues can be linked to the low occurrence of the interpersonal consequence value (CSQ INT), which relates to the "turn-yielding" (Crible & Degand, 2019) characteristic of some ending presence of 'so', as in the following example from *Love is Blind*:

(10) That's one of the things I love about her, **so**...

When 'so' is used at the end of a sentence as an abstract consequence marker, it hints to the interlocutor that the argument is over and that they can take their turn. The turn-yielding function of 'so' leaves the consequence link to be restored by the interlocutor.

It also appears that the monitoring value (MNT) is only present in film dialogues, as the hedging one (HDG) belongs here in the reality TV show. As a comparison, some examples of monitoring and hedging from the corpora are presented here. From the film *Fight Club* in the assembled corpus, the usage of 'so' as interpersonal monitoring function (MNT INT) can be found:

(11) …I know we' re all -- we' re all dying, all right? But you' re not dying the way Chloe back there is dying.
-**So**?
-So you' re a tourist.

It can be observed that ‘so’ is used here to maintain the flow of the dialogue by asking the interlocutor to continue their point. To contrast this usage with the other form of monitoring present in the assembled corpus, here is an example of sequential monitoring (MNT SEQ) from *Avatar*:

(12) -Plus it 'll help to keep you sane for the next six years.
-All right.
-Whatever.
-**So**…
-Well, here I am, doing science.

The function of ‘so’ is to control the discussion by stagnation. However, there is no apparent link with what has been stated earlier.

The interpersonal hedging function (HDG INT) of ‘so’ from the LiB corpus (episode 7) is illustrated here:

(13) …he's just always wanted our kids to marry
within our race. **So**... I don't know how he'll handle that.

The use of ‘so’ has a face-saving function in the above sentence (13). What the speaker said is politically on edge, and therefore ‘so’ appears to mitigate the tension.

Film dialogues are written by experienced scriptwriters and edited several times beforehand. Therefore, it seems that ‘so’ is mainly used to control the flow of discussions. However, in the reality TV show, ‘so’ is used to put some relativity and hesitation rather than controlling the speech.

‘So’ is also used with a topic value (TOP), where there are no semantic links with what has been said before, as illustrated by the following example from the LiB corpus (episode 1):

(14) -Hello!
-Hi! **So**, what are some of your biggest turn-offs?

The function of ‘so’ here is to begin a topic, as nothing has been said before except for the salutations. This topic usage of ‘so’ is overused in high proportion in film dialogues compared to reality TV show.

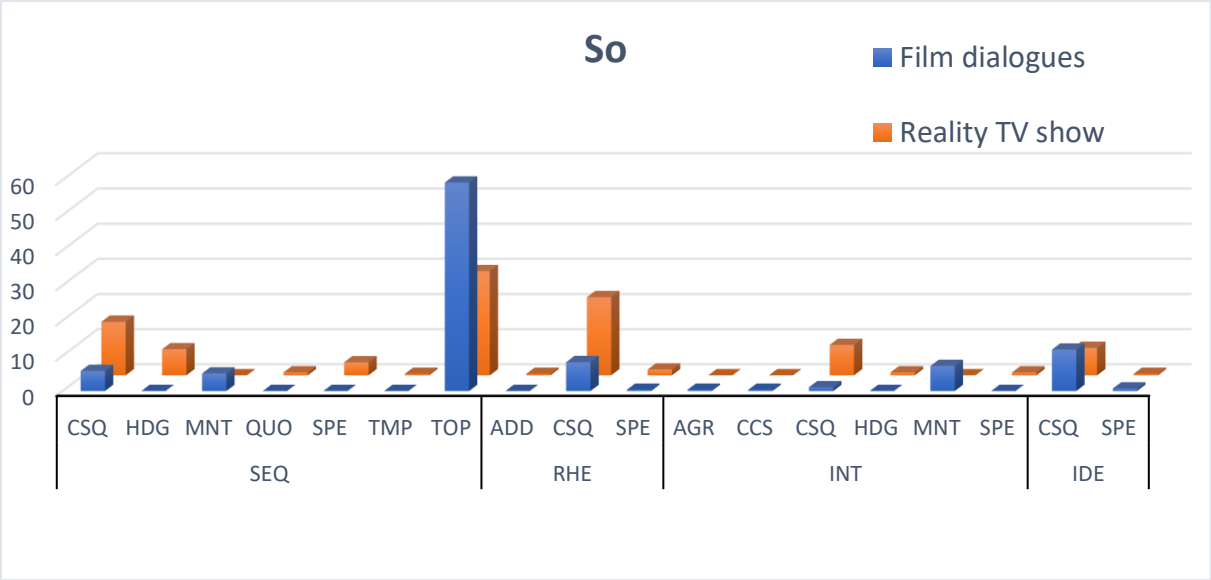


Figure 9. Semantic and pragmatic values of 'so' in film dialogues and reality TV show.

4.1.3 WELL

'Well' appears to occur mainly in the first position of a sentence, as shown in Figure 10. It is also evident that 'well' rarely appears at the end of a sentence. The results for film dialogues and reality TV show are similar; no difference can be accounted regarding the position.

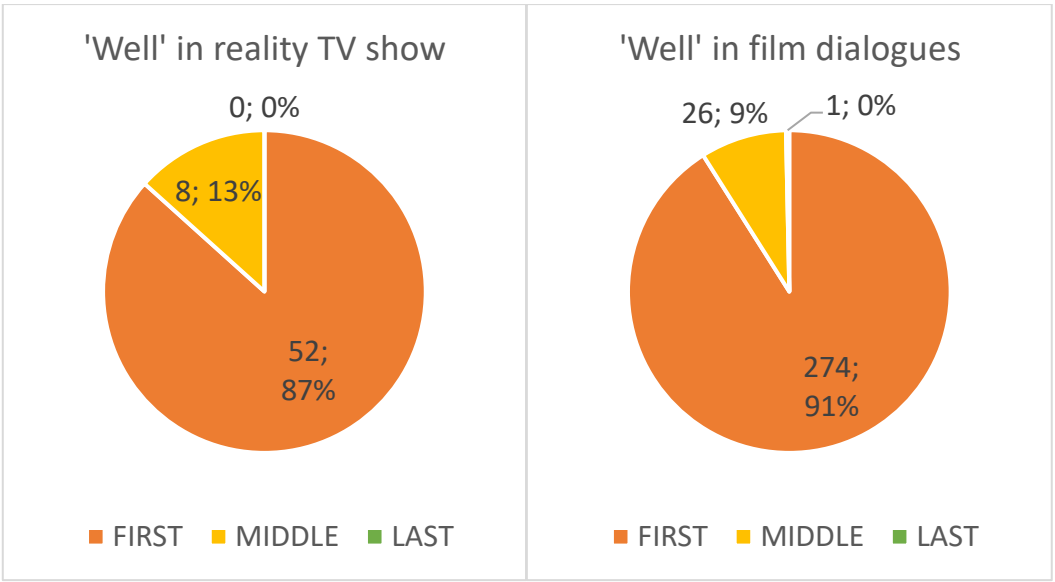


Figure 10. Position of 'well' in reality TV show and film dialogues.

An interesting difference can be made between the two corpora regarding the frequency of ‘well’. It is the most common discourse marker in film dialogues but is the least frequent one (considering the three DMs selected for this study in particular) for the reality TV show. Aijmer and Simon-Vandenbergen (2003) have analysed ‘well’ and its Dutch counterpart ‘nou’. They state that ‘well’ appears in a more formal context. So, one explanation can point to the register, as reality TV shows are less formal, while film dialogues can have a broader range of register levels. The variations of English could also be an element. Some film characters in the assembled corpora use British English, while all the participants in the LiB corpus are Americans.

Considering the semantic and pragmatic values of ‘well’, some differences can be noted, as shown in Figure 11. A similar phenomenon concerning the hedging (HDG) and monitoring (MNT) values of ‘so’ can also be observed for ‘well’. The hedging value appears mainly in a reality TV show, whereas monitoring seems to be more frequent in film dialogues. The following example coming from the assembled corpus illustrates this fact (film: *Bridge of Spies*):

(15) If that is what you're saying, **well**, then there is never any limit to our liability and that is the end of the insurance business.

This occurrence of ‘well’ in (15) has been considered as a monitoring function. The aim of this discourse marker is not to stall for time but to control the flow of the discourse. In the LiB corpus, however, ‘well’ is used for hedging (an example from episode 1):

(16) -My dog sleeps with me every night, so… I love that.
-Mine, too. **Well**, I'm just... I'm a big cuddler.

The function of ‘well’ is to fill a gap originating from hesitation. This is highlighted by the following sentence, where a false start appears – “I’m just…” confirming the hesitation.

The rhetorical specification (SPE RHE) is dominant and specific to film dialogues, as illustrated by the following example (17) from the film *Touching the void*:

(17) And he was lowering me on a 9mm, **well** 8.8mm rope. That's that thick.

‘Well’ is used in this case as a specification marker, adding details to the first statement. According to the data, the rhetorical specification function of ‘well’ is rarely used in the reality TV show. The discourse marker is also used for changing topics (TOP), distributed equally in film dialogues and reality TV show (see Figure 10).

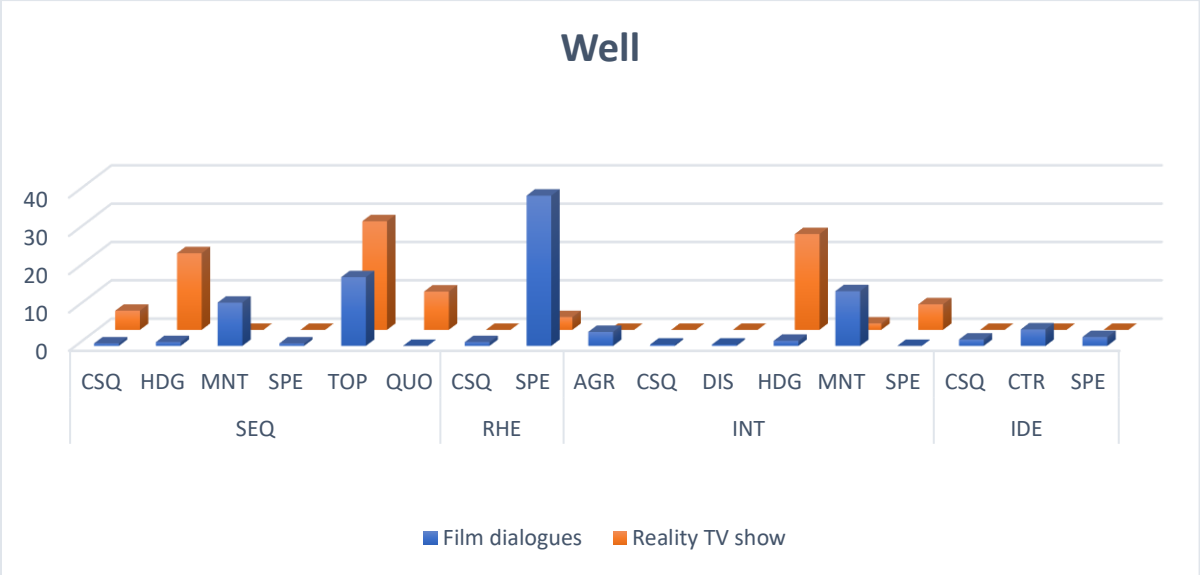


Figure 11. Semantic and pragmatic values of 'well' in film dialogues and reality TV show.

4.2 Film dialogues and Reality TV show

Film dialogues are written to be spoken and look natural. However, the preparation is still a bias. To contrast, reality TV shows are more natural, with long, unprepared dialogues, hesitations, and repetitions. What can be observed from the comparative study is that positions are similar, except for a few cases. ‘You know’ at the end of a sentence seems to be underused in the reality TV show compared to film dialogues. The usage of ‘you know’ in a natural, less formal conversation from the reality TV show appears to be less literal and more incidental in the middle of sentences. ‘So’ expressing the interpersonal consequence at the end of a sentence is underused in film dialogues. The usage of ‘well’ and ‘so’ are also different, mainly as monitoring in film dialogues and as hedging in the reality TV show. The criterion for the preparation of the dialogues is of particular importance here. ‘Well’ is overused in film dialogues and appears to have a broader variety of functions than in the reality TV show. ‘You know’ and ‘so’, however, have more variations in reality TV show than in film dialogues.

Using the three discourse markers shows some notifiable differences in whether the dialogues are scripted or not. The register may also influence their usage. The reality TV show presents a more informal context, while the film dialogues have broader possibilities. Further in-depth analysis is necessary to assess whether reality TV show is more similar to natural speech than film dialogues. This work is but a first step in that direction.

4.3 Contrastive analysis

The second part of this study focuses on the translated subtitles of the former mentioned discourse markers into their French and Dutch counterparts. As stated before, subtitles must follow some rules, and this bias influences the total number of discourse markers translated. The assembled corpus and the LiB corpus have been combined for this analysis, as the translation process into subtitles does not differ for the film dialogues or reality TV show.

As shown in Figure 12, only 29% of the discourse markers are translated into French and only 37% into Dutch.

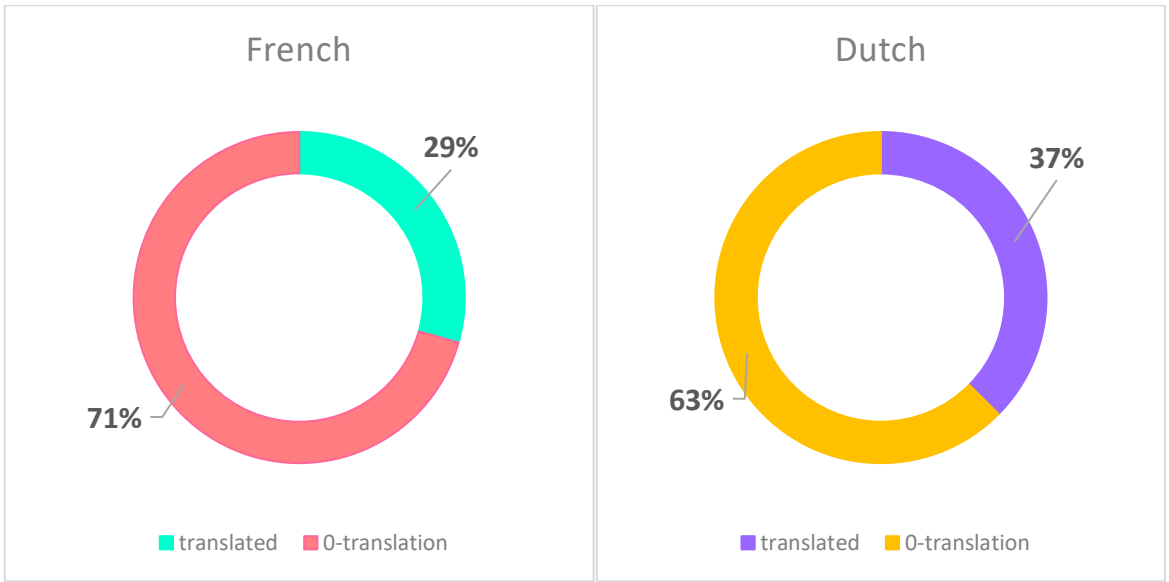


Figure 12. Percentage of DMs translated in French and Dutch.

Of the three DMs considered, ‘so’ is most likely to be translated. From the 523 occurrences in the two corpora, 48% are translated into French and 51% into Dutch.

‘You know’, however, is the most unlikely to have a translation, with more than a 70% chance of being lost in translation out of the 236 occurrences. ‘Well’ follows the same pattern, with a few more cases translated into Dutch than ‘you know’. Out of the 361 occurrences of ‘well’, more than 60% are lost in translation (see Figure 13).

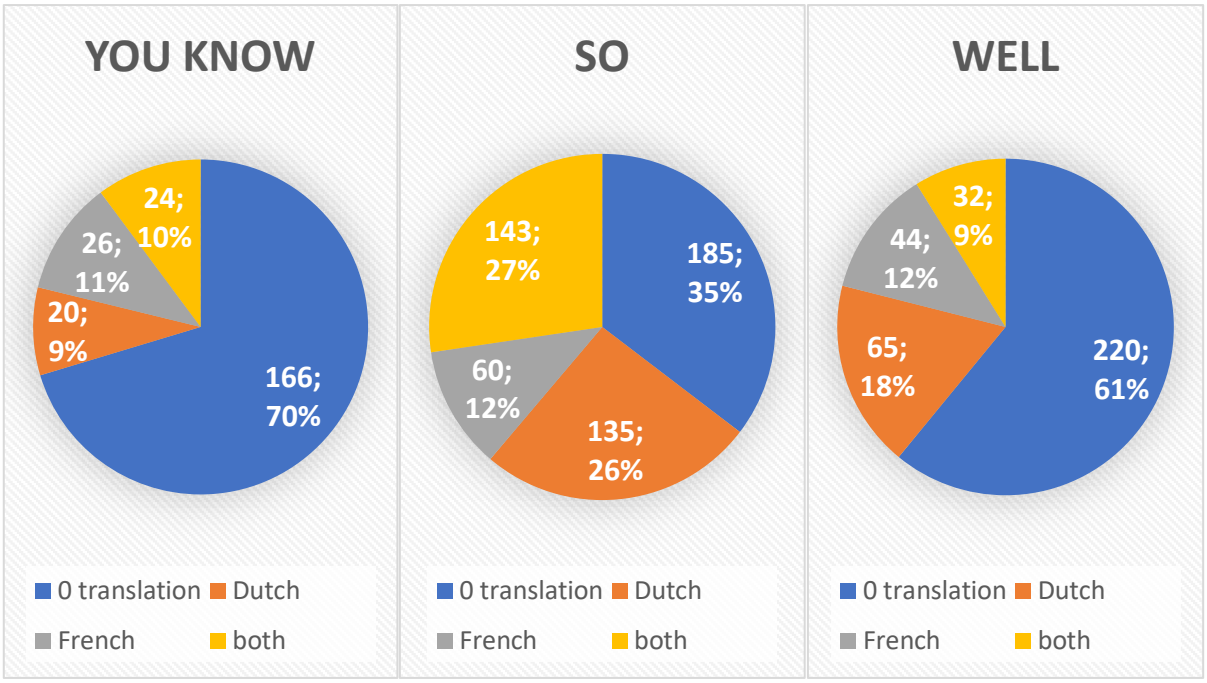


Figure 13. Detailed results of translations for the three DMs.

When combined, 51% of the 1120 discourse markers considered for this study have not been translated. Additionally, 20% are not translated into French, and 11% are not translated into Dutch. Only 18% are translated into both languages, as shown in Figure 14. These results are in close agreement with previous studies (Degand, 2015; Connors, 2016), which suggests that more than 50% of the DMs were lost in translation.

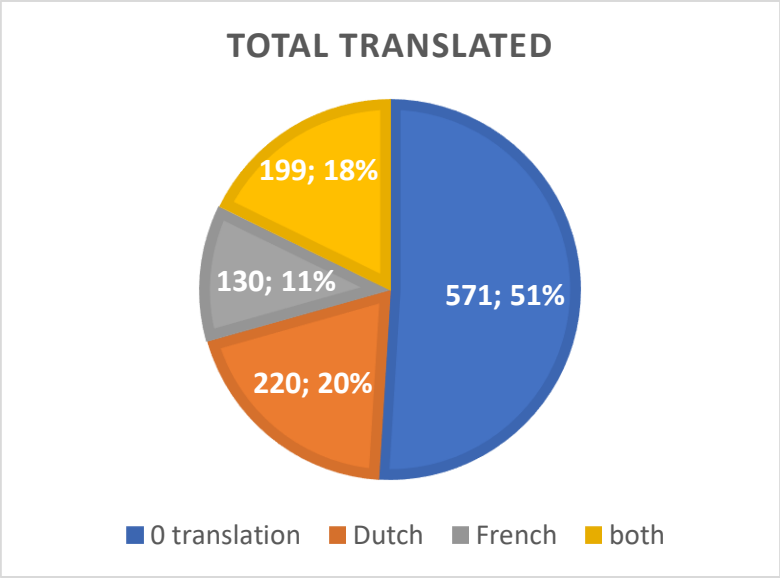


Figure 14. General results of translation for the three DMs.

4.3.1 Semantic and pragmatic values

The analysis of the semantic and pragmatic values aims to assess if any patterns or tendencies can be identified for these discourse markers. The four pragmatic values, ideational, rhetoric, sequential and interpersonal, are organised from the most concrete to the most abstract (see Section 4.1). As illustrated in Table 4, the distribution is unequal. It seems, however, that the ideational domain is the one which is most likely to be translated, with approximately 50% of the occurrences being translated.

Table 4. Distribution of the pragmatic values among the three DMs.

	OCCURRENCES	ZERO TRANSLATION	ONLY IN FRENCH	ONLY IN DUTCH	BOTH TRANSLATIONS
IDE	**84**	28	6	15	**35**
RHE	206	116	21	44	25
SEQ	589	316	71	110	92
INT	241	111	32	51	47

As far as the semantic functions are concerned, consequence (CSQ) is the one which is most likely to have a translation. On the contrary, about 50% of the time, specification (SPE) and monitoring (MNT) are lost in translation. Topic (TOP) follows the average of 50%. The distribution is, however, still unequal (see Table 5). Following the trend discussed in the previous paragraph, the ideational consequence (CSQ IDE) appears to have a high probability of being translated. On the other hand, sequential (MNT

SEQ) and interpersonal monitoring (MNT INT) and rhetorical specification (SPE RHE) have the least chance of being retained in translated subtitles. Sequential (HDG SEQ) and interpersonal hedging (HDG INT) are implicitly conveyed rather than translated.

Table 5. Assembled results of semantic and pragmatic values for the three DMs.

		ZERO TRANSLATION	ONLY IN FRENCH	ONLY IN DUTCH	BOTH TRANSLATIONS	TOTAL
ADD	RHE	0	0	0	1	1
AGR	INT	4	0	6	2	12
CCS	INT	0	0	1	0	1
CSQ	IDE	11	5	10	**31**	57
CSQ	RHE	22	10	28	20	80
CSQ	SEQ	17	3	19	19	58
CSQ	INT	6	2	9	8	25
CTR	IDE	8	0	3	2	13
DIS	INT	1	0	0	0	1
HDG	SEQ	14	2	9	10	35
HDG	INT	17	3	7	3	30
MNT	SEQ	**94**	17	12	**10**	133
MNT	INT	**80**	26	26	34	166
QUO	SEQ	6	2	2	0	10
SPE	IDE	9	1	2	2	14
SPE	RHE	**94**	11	16	**4**	125
SPE	SEQ	9	2	4	1	16
SPE	INT	3	1	2	0	6
TMP	SEQ	0	0	0	1	1
TOP	SEQ	**176**	45	64	51	336
	Total	571	130	220	199	1120

In summary, pragmatic ideational values seem more likely to be translated. The less abstract the relation is, the more chances it will be translated. As for the semantic values, the consequence values are more likely to be retained. As the consequence value is mainly represented by ‘so’ (see Section 4.1.2, Figure 9), it matches the results of ‘so’ being the discourse marker with the most chances to be kept in the final translation.

4.3.2 Translations in French and Dutch

This section will analyse the equivalent DMs of ‘you know’, ‘so’, and ‘well’ adopted while translating into French and Dutch. The results are presented in histograms from the assembled corpus and the LiB corpus separately to see if any major differences can be observed in terms of translation choices.

4.3.2.1 YOU KNOW

The variation of the equivalent translational phrases for ‘you know’ in French and Dutch is presented in Figure 15. The literal translation of ‘you know’ is ‘tu sais’ or ‘vous savez’ in French and is the most preferred choice in film dialogues (adding to that ‘tu sais bien’). In the reality TV show, however, the translator prefers the alternative ‘tu vois’ (you see) and its variations i.e., ‘tu vois ce que je veux dire’ and ‘vous voyez’. The usage of the polite or plural ‘vous’ pronoun is naturally less present in the reality TV show, due to the informal context of dating. It can also be noted that there are more variations for translated ‘you know’ in the reality TV show. Among others, a significant number of ‘bon’ and the presence of ‘du genre’ (an approximant translation of the adverb ‘like’) occur frequently. ‘Bon’ hints at the higher presence of the hedging and monitoring function of ‘you know’ in the LiB corpus, while ‘du genre’ stands for the quoting or specification function, the latter illustrated by the following example out of the LiB corpus (episode 1):

(18)	All these girls love my jokes, every single one of them. **You know**, Oh, next one laughed."	Les filles ont aimé mes blagues, toutes mes blagues. **Du genre** : Oh, elle a ri.

The presence of the semi-colon in the French subtitles hints at the ideational specification role of the discourse marker.

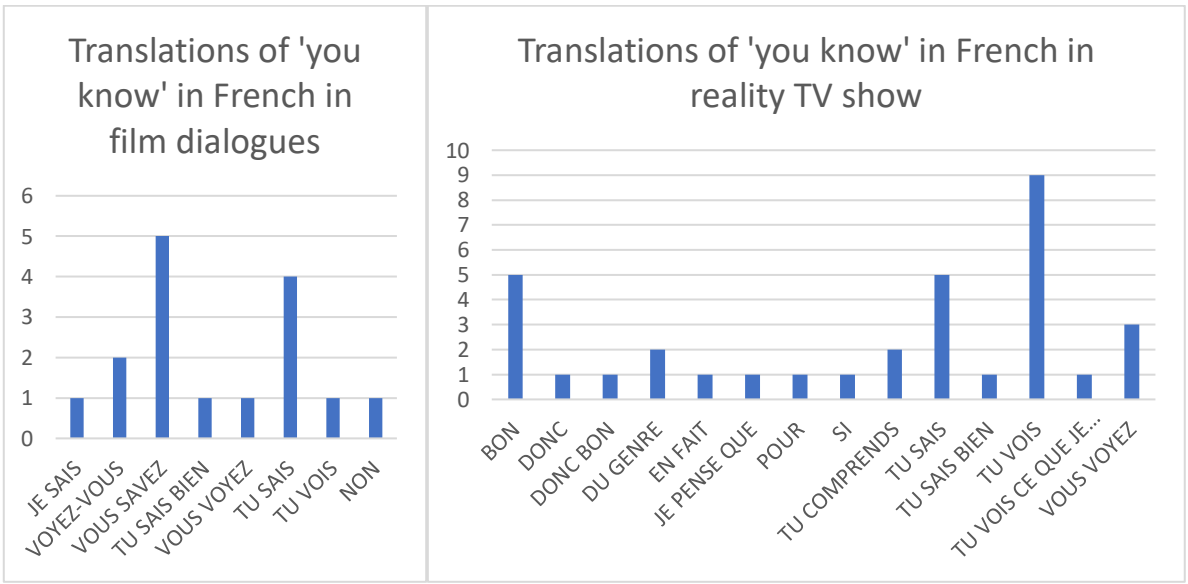


Figure 15. Translation choices for 'you know' in French in film dialogues and reality TV show.

The results for the Dutch choices follow the same pattern. The literal translation ‘weet je’ is dominant in film dialogues with all its variations: ‘je weet wel’, ‘weet jij het’, ‘weet u’ and ‘weet je wel’. The Dutch language offers more possible variations than in French. The possibility exists in Dutch, to begin with, the pronoun (‘je weet’) or with the verb (‘weet je’), which is not possible either in English or French. There are also different pronouns for the translation of ‘you’, i.e., ‘je’, ‘jij’ and ‘u’ (the plural form ‘jullie’ and the less standard forms ‘ge’ and ‘gij’ are not represented in the corpora). ‘Weet je’ and some variations are also dominant in the reality TV show (excluding the polite ‘u’ pronoun due to the special context). ‘Snap je’, an informal variation of ‘do you understand’ (or ‘do you get it’), is also a strong choice, probably because of the informal dating context (see Figure 16).

Figure 16. Translation choices for 'you know' in Dutch in film dialogues and reality TV show.

4.3.2.2 SO

‘So’ is the most translated discourse marker among the three selected ones. As shown in Figure 17, the literal translation ‘alors’ is dominant in film dialogues (and its variation ‘et alors’). ‘Donc’ (~thus) is also present but in a minor proportion. On the contrary, in the LiB corpus, ‘donc’ is the primary choice. ‘Donc’ can imply a consequence relation as illustrated in the following example from the LiB corpus:

(19)	See, I don't live by myself, **so** for me it's the car.	Je ne vis pas seule, **donc** je chante dans la voiture.

The consequence of not living by herself is that the speaker prefers to sing in her car, with ‘so’ being translated as ‘donc’ in French. The preference for ‘alors’, a literal translation, or ‘donc’ could be explained by the register or by the more substantial effect of ‘translationese’ in film dialogues. Furthermore, the numbers of occurrences of ‘bon’ differ entirely. ‘Bon’ is scarcely present in film dialogues but is the second choice for the reality TV show (also mixed with another DM: ‘bon alors’ and ‘donc bon’). The presence of the hedging ‘so’ in the reality TV show can explain this difference, as illustrated by the following example from the LiB corpus:

(20)

-It's just, there's a physical discomfort still -in kind of moving into this phase. -Right, that's okay. **So**... I know she's dated taller guys before, and Mark is shorter.	- J'y arrive, petit à petit. - C'est bien. **Bon**, Jessica est sortie avec des hommes plus grands, et Mark est plus petit.

The hedging function of ‘so’ is here hinted by the suspension dots, where the speaker uses ‘so’ to stall.

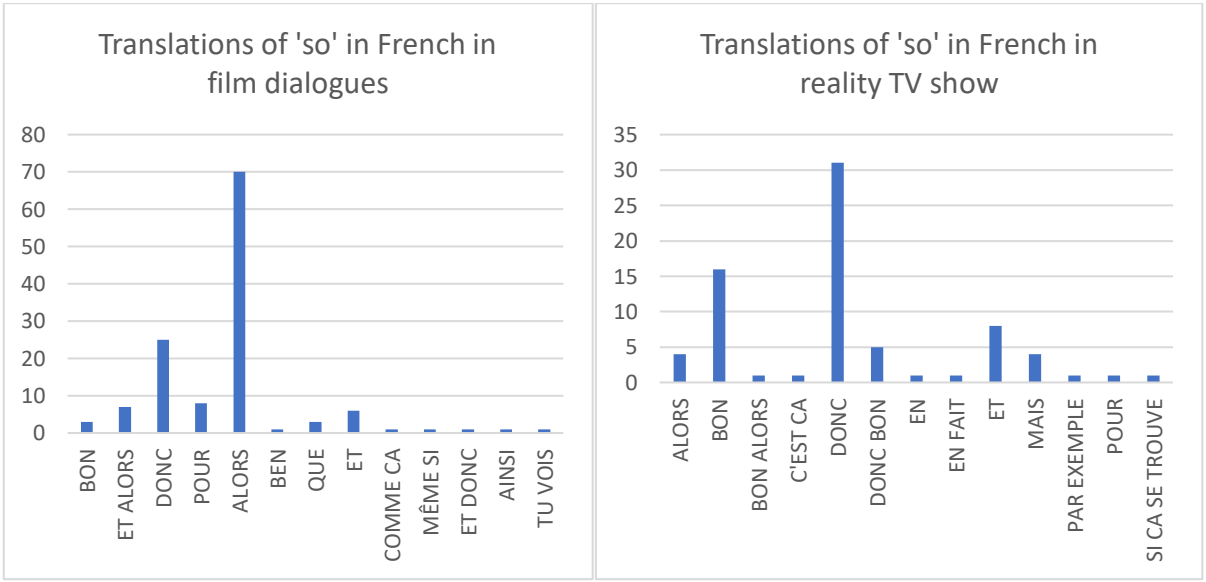


Figure 17. Translation choices for 'so' in French in film dialogues and reality TV show.

As illustrated in Figure 18, the results for the translations of ‘so’ in Dutch are less disputed. ‘Dus’ (~thus) is massively chosen. The literal translation ‘dan’ is also present but mostly incidental. A correlation can be made with the presence of ‘donc’ in French, ‘dus’ being a possible translation for ‘donc’. ‘Dus’ conveys, in the same way as ‘donc’, a consequence link, for example from the LiB corpus:

(21)	I'm nervous, but I like talking, **so**...	Ik ben nerveus, maar ik praat graag, **dus**...

This example demonstrates an interpersonal consequence function of ‘so’. The discourse marker ends the argument, hinting to the interlocutor that they can take turns. The consequence link is left to the interlocutor to be reconstructed.

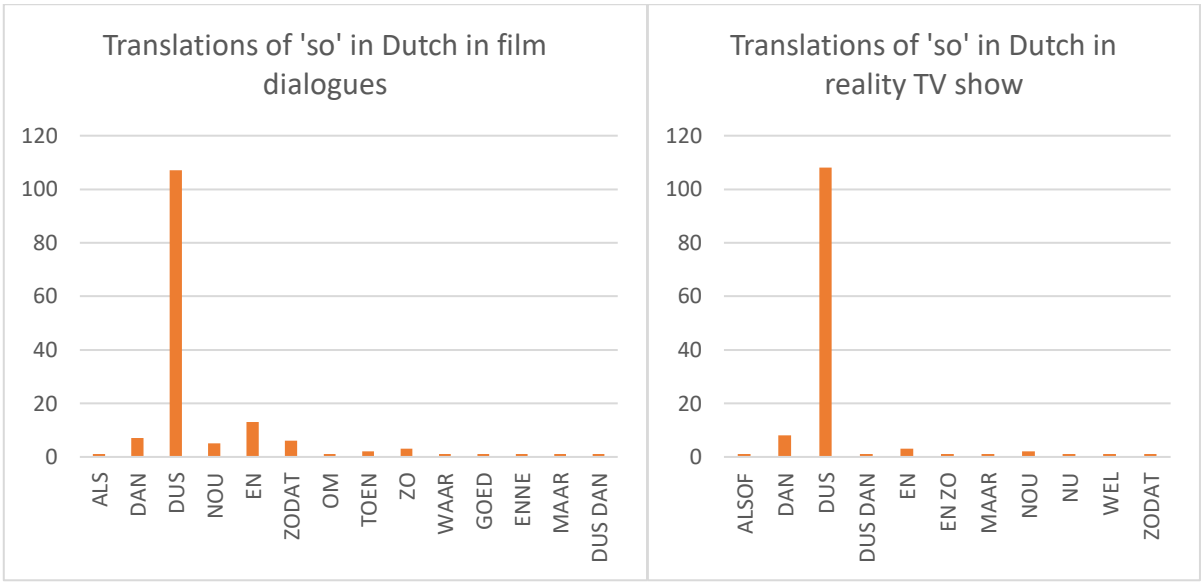


Figure 18. Translation choices for 'so' in Dutch in film dialogues and reality TV show.

4.3.2.3 WELL

As illustrated in Figure 19, the literal translation of ‘well’ in French, ‘bien’ is again dominant in film dialogues with a variation of it, ‘eh bien’ being the first choice. However, this point needs to be nuanced. In the film *Spectre*, the translator preferred to use ‘eh bien’ (about 14 times out of 31 occurrences) for ‘well’. Eventually, this has falsely placed ‘eh bien’ as the most preferred choice of translation for ‘well’ with a maximum of 21 occurrences. It means that 14 of the 21 ‘eh bien’ emanated from a single film translated by one translator. The hypothesis is that the translation quality is not always equivalent. It nuanced the dominance of ‘eh bien’ but does not invalidate its presence; even by removing the 14 occurrences of the film *Spectre*, ‘eh bien’ remains a valid choice. Nevertheless, ‘bien’ is also a strong choice, as well as ‘alors’ (so) and ‘bon’ (good). The broad range of translation choices can be linked with the more diverse semantic values established in the assembled corpus than in the LiB corpus.

Regarding the reality TV show, ‘bon’, ‘eh bien’ and ‘alors’ are the first choices, which follows the same trend as the results for the film dialogues.

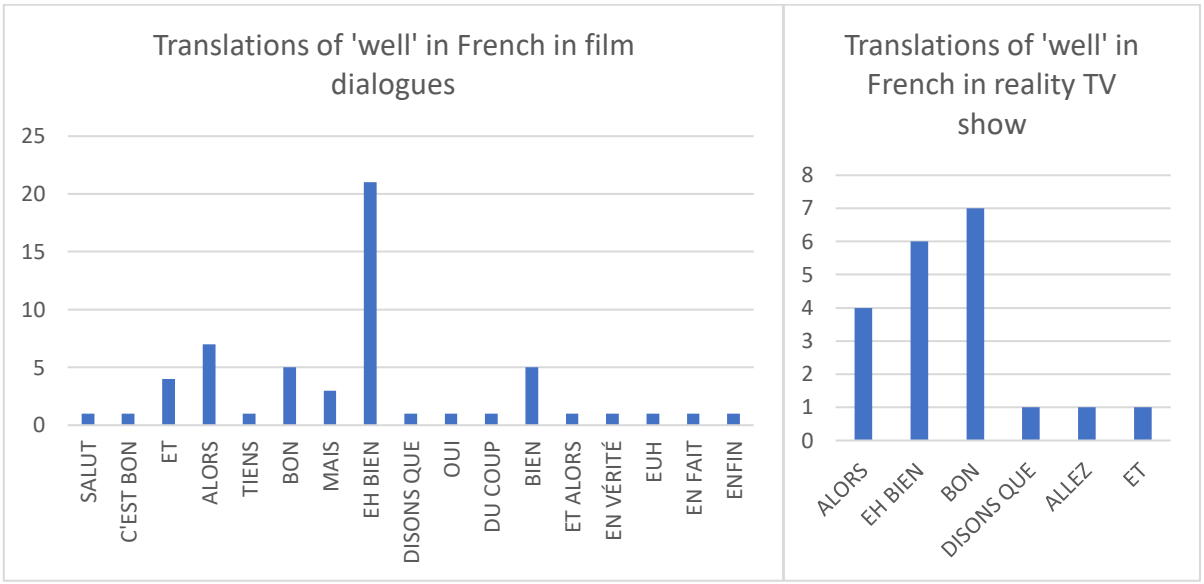


Figure 19. Translation choices for 'well' in French in film dialogues and reality TV show.

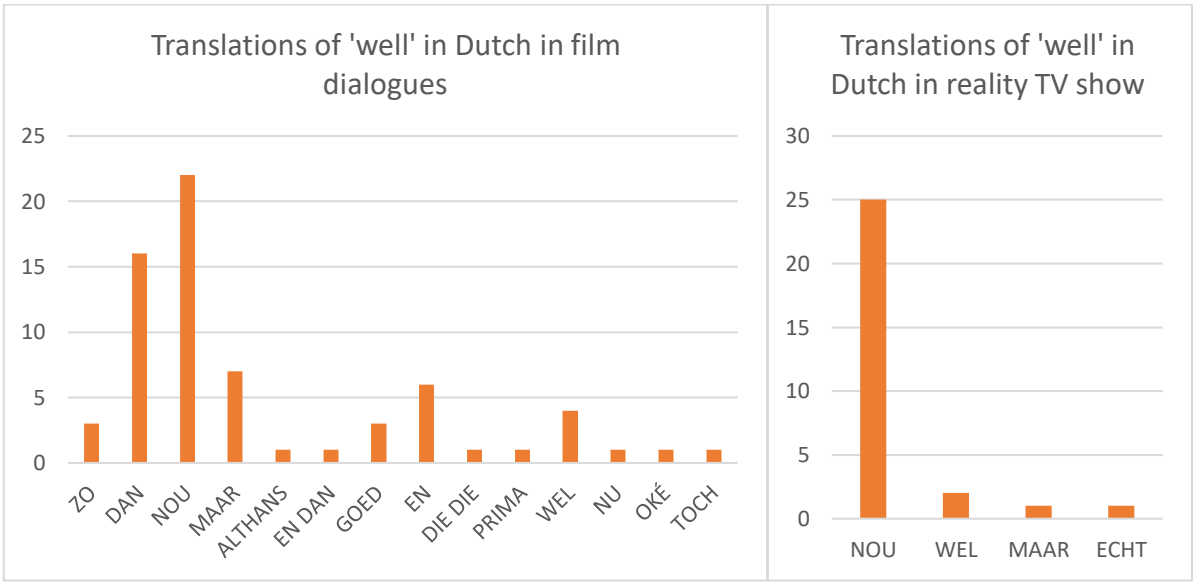


Figure 20. Translation choices for 'well' in Dutch in film dialogues and reality TV show.

As for Dutch, ‘nou’ is the dominant choice for translating ‘well’ (see Figure 20). ‘Dan’ (so) and with less importance ‘maar’ (but) and ‘en’ (and) are also represented translation choices. The presence of more diverse semantic values in film dialogues explains the broader choice of DMs. ‘Nou’ is also dominant in the reality TV show, but the other choices are anecdotal. Finally, the literal translation ‘wel’ is present in both corpora but does not appear to be a strong choice. In conclusion, the literal translation

‘wel’ is not convincing as an alternative for ‘well’, probably not conveying the same semantic values or being used in a different register level. In order to attest to this, further analysis is necessary and is beyond the scope of this present study.

5 Discussion

There are a few limitations to the present study that need to be discussed. The aim is to study discourse markers in real-time conversations, and for that, film and reality TV dialogues were selected. This choice was also driven by the contrastive part of the analysis, considering French and Dutch subtitles. Other corpora could have been used, as films and reality TV both present some biases. However, every speech corpus does present biases, as participants need to be aware that they are recorded. In that case, they would often adapt their speech accordingly. Film dialogues are scripted but in an attempt to recreate or mimic real speech. This allows the scriptwriter to think and rewrite the dialogues, even if the aim is to be as close as possible to natural conversation. The dialogues in reality TV are not scripted, but the context and the editing are also two factors of bias. First, some scenes may have been recast several times, hindering the natural part of the speech. Then, the show is a product of all the edited emotional scenes, with no guarantee that they were edited in a chronological order. This, however, does not invalidate film and reality TV dialogues as a source for oral corpora. A bias-free corpus has not yet been achieved.

Furthermore, the choice of discourse markers is another limitation. The three discourse markers in this study were selected due to their high frequency in the film captions. Even if a broader selection would present more precise results, this study paves the way for further investigations.

The contrastive analysis was conducted using the annotation system developed by Crible and Degand (2019) and the position of the DM. Other criteria, such as the level of formality and information about the context could also add some light to the results and some precision to the choice of translation in French and Dutch. The different DMs do not present the same characteristics in different languages and these criteria may hint to more explanations.

Given the time and the resources, the double-coding was done by the author alone. However, involving more coders would allow more objectivity to this study. Its replication then appears to be interesting to validate the results.

6 Conclusion

This study investigated the differences between the use of discourse markers in film and reality TV dialogues. Film dialogues are written beforehand and recited by actors. The dialogue writers had time to think, organise and rewrite the scripts. In reality TV shows, there supposed to be no script or lines, even if some may be more scenarised than others. In this study, the reality TV show '*Love is Blind'* was selected as the reference as it does not involve votes from the viewers deciding the fate of the participant's time in the event or any sort of public participation. Therefore, participants have more liberty to speak freely without any external constraints compared to other popular reality TV shows. To conduct the study, the annotation system developed by Crible and Degand (2019) was used, and the position of the DMs was also considered. Three DMs were selected: 'you know', 'well' and 'so' as these were the most common in the film dialogues.

At this point, it seems wise to refresh the research questions in the mind. The three research questions and the secondary questions were:

1. How does the use of discourse markers differ between film dialogues and reality TV shows?
2. To what extent are discourse markers translated in subtitles?
3. What factors influence the translation choices of DMs in subtitles?
    a. Can a contrastive analysis of the translations of certain DMs reveal common patterns among different languages and shed light on the translator's choices?

The comparison reveals that the usage of discourse markers does differ between the two sources. The pragmatic domains of the three discourse markers does not offer much variation, the sequential domain remains the most used domain. However, a slight difference can be noted as the interpersonal domain is more present in the reality TV show than in the film dialogues. The semantic functions present a major difference, as specification is massively used in film dialogues in comparison with reality TV show, monitoring is also more present in films, hedging being used mainly in reality TV show.

As far as ‘you know’ is concerned, its position seems to be particularly controlled in the film dialogues, and the proportion appears to be equally distributed. However, in the reality TV show, about 50% of the time, the participants use it in the middle of the phrase. ‘You know’ is also used more for its sequential domain in the reality TV show, while the interpersonal proportion is higher in film dialogues. Monitoring is unsurprisingly the main function of ‘you know’.

‘So’ is mainly used in first position for both films and reality TV show. However, ‘so’ also appears in last position (11%) but only in reality TV show, as an interpersonal consequence marker. The sequential domain is still dominant, with more variation in the reality TV show. The main functions of ‘so’ appears to be consequence and topic.

The relative proportion of ‘well’ is completely different in films and reality TV dialogues. Only 12% of the discourse markers (60 out of 470) are ‘well’ in the reality TV show. Whereas 46% of the DMs appearing in 37 films are ‘well’ (301 out of 650). The completely different proportion can be explained by the register. A broader range of speech levels is present in films, where the context can vary significantly. In the reality TV show, the level of formality stays rather low, as in most of the episodes, the participants are dating. The occurrence position of ‘well’ is similar in both media, mainly in first position, but also sometimes appearing in the middle. The preferred domain is sequential for both sources, but the semantic functions vary. Specification seems to be the main use in film dialogues, whereas hedging is the most used function in the reality TV show.

The second part of the study focuses on the translations or the absence of translations of the three selected discourse markers in French and Dutch. Only 29% of the 1120 discourse markers are translated into French and 37% into Dutch. When combined, 51% of the discourse markers are not translated into French or into Dutch. This confirms the hypotheses drawn from previous studies (Degand, 2015; Connors, 2016).

A clear pattern is however difficult to identify as a reason for being translated or not. From the results, it appears that the ideational domain is most likely to be translated. According to Abuczki et al. (2018) with their study on ‘and’, the sequential domain was less likely to be translated. However, in this study, the three other domains, rhetorical, sequential, and interpersonal were close with around 70% chance of being lost in translation. The consequence relation has the highest chance of being translated

among the semantic functions of the three discourse markers selected in this study. As ‘so’ is the clear representation of the consequence link, it is also the most translated DM out of the three.

For the three discourse markers, some translations appear to be more common and preferred than others. ‘You know’ is mostly translated by its counterparts, ‘tu sais’ or ‘tu vois’ in French (and their nuances) and ‘weet je’ or ‘weet u’ (and their nuances) in Dutch. Even if French and Dutch offer more possible nuances, the direct translation seems to apply in the same context. The use of ‘tu vois’ (‘you see’) in French is however a particularity, as it appears to also be a good alternative for you know. The Dutch phrase ‘snap je’ (‘do you get it’) appears to be a good alternative in the reality TV show, probably in a less formal context. A literal translation of ‘so’, ‘donc’ in French and ‘dus’ in Dutch seems to be the main choice of translations. However, ‘alors’ is preferred to ‘donc’ in film dialogues. ‘Nou’ is by far the first choice of translation of ‘well’ in Dutch, even if ‘wel’ exists in the Dutch language, it does not seem to be a strong choice. ‘Eh bien’ or ‘bon’ can be considered as some literal translation of ‘well’ and are the first choices in French. Therefore, it can be concluded that there are some patterns for the translation of these three discourse markers, where one translation is clearly preferred. Some literal translations always seem to be a strong pick, except for the Dutch ‘wel’, where ‘nou’ is by far the preference.

References

Abuczki, Á., Burkšaitienė, N., Crible, L., Nedoluzhko, A., Furkó, P., Valūnaitė Oleškevičienė, G., ... & Zikánová, Š. (2018). Translation of “and” in a parallel TED Talk corpus of English, Czech, Hungarian, Lithuanian and French: functions and omissions. In *TextLink2018--Final Action Conference* (p. 4).

Aijmer, K., & Simon-Vandenbergen, A. M. (2003). The discourse particle well and its equivalents in Swedish and Dutch. *Linguistics, 41*(6), 1123–1161.

Beeching, K. (2016). *Pragmatic markers in British English: Meaning in social interaction*. Cambridge University Press.

Biagini, M. (2010). Les sous-titres en interaction: le cas des marqueurs discursifs dans des dialogues filmiques sous-titrés. *Glottopol*, *15*, 18-33.

Bolden, G. B. (2009). Implementing incipient actions: The discourse marker ‘so’ in English conversation. *Journal of pragmatics*, *41*(5), 974-998.

Bosker, H. R., Badaya, E., & Corley, M. (2021). Discourse markers activate their, like, cohort competitors. *Discourse Processes*, *58*(9), 837-851.

Buysse, L. (2012). So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics*, *44*(13), 1764-1782.

Buysse, L. (2014). “So what's a year in a lifetime so.” Non-prefatory use of so in native and learner English. *Text & Talk*, *34*(1), 23-47.

Buysse, L. (2017). The pragmatic marker you know in learner Englishes. *Journal of Pragmatics*, *121*, 40-57.

Chaume, F. (2004). Discourse markers in audiovisual translating. *Meta*, *49*(4), 843-855.

Connors, M. D. (2016). The pragmatic particles enfin and écoute in French Film and TV Dialogue.

Crible, L., Abuczki, Á., Burkšaitienė, N., Furkó, P., Nedoluzhko, A., Rackevičienė, S., ... & Zikánová, Š. (2019). Functions and translations of discourse markers in TED Talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, *142*, 139-155.

Crible, L., & Degand, L. (2019). Domains and functions: A two-dimensional account of discourse markers. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Cuenca, M. J. (2008). Pragmatic markers in contrast: The case of well. *Journal of Pragmatics, 40*(8), 1373–1391.

Degand, L. (2015, May 06-09). *Marqueurs discursifs en traduction à l'oral. Les sous-titres de film sont-ils le reflet de la conversation spontanée ?* 4th International Symposium of Romance Discourse Markers, Heidelberg, Germany.

Diễm, P. N. (2023). The Function of Discourse Marker "WELL" In English and Vietnamese Communication. *South Asian Res J Human Soc Sci*, *5*(2), 25-33.

Fox Tree, J. E. (2010). Discourse markers across speakers and settings. *Language and linguistics compass*, *4*(5), 269-281.

Fox Tree, J. E., & Schrock, J. C. (2002). Basic meanings of you know and I mean. *Journal of Pragmatics*, *34*(6), 727-747.

Fraser, B. (1996). Pragmatic markers. *Pragmatics*, *6*, 167-190.

Fraser, B. (1999). What are discourse markers?. *Journal of pragmatics*, *31*(7), 931-952.

Furkó, B. P. (2014). Perspectives on the Translation of Discourse Markers. *Acta Universitatis Sapientiae, Philologica*, *6*(2), 181-196.

Heine, B. (2013). On discourse markers: Grammaticalization, pragmaticalization, or something else? *Linguistics, 51*(6), 1205–1247.

Heine, B., Kaltenböck, G., Kuteva, T., & Long, H. (2021). *On the rise of discourse markers* (Vol. 219, p. 23). John Benjamins Publishing Company.

Hosogoshi, K. (2016). Effects of Captions and Subtitles on the Listening Process: Insights from EFL Learners' Listening Strategies. *Jalt Call Journal*, *12*(3), 153-178.

Huang, L. F. (2018). A Corpus-Based Exploration of the Discourse Marker Well in Spoken Interlanguage. *Language and Speech, 62*(3), 570-593.

Innes, B. (2010). “Well, that’s why I asked the question sir”: Well as a discourse marker in court. *Language in society*, *39*(1), 95-117.

Karankata, A., Negri, M., & Turchi, M. (2020). MuST-Cinema: a Speech-to-Subtitles corpus. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 3727–3734. https://aclanthology.org/2020.lrec-1.460

Levshina, N. (2016). Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast, 16*(1), 84-117.

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System*, *63*, 134-148.

Prasad, R., Webber, B., & Lee, A. (2018, August). Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (pp. 87-97).

Pujadas, G., & Muñoz, C. (2020). EXAMINING ADOLESCENT EFL LEARNERS’TV VIEWING COMPREHENSION THROUGH CAPTIONS AND SUBTITLES. *Studies in Second Language Acquisition*, *42*(3), 551-575.

Ro, E., & Jung, H. (2022). “Aratda!”: Intersubjectivity-in-action in a multilingual Korean reality TV show. *Journal of Pragmatics*, *201*, 89-117.

Rouchota, V. (1996). Discourse connectives: What do they link? *UCL Working Papers in Linguistics*, *8*, 1–15.

Roze, C., Danlos, L., & Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours. Revue de Linguistique, Psycholinguistique et Informatique. A Journal of Linguistics, Psycholinguistics and Computational Linguistics*, *10*, 10. https://doi.org/10.4000/discours.8645

Rysová, M. (2017). Discourse connectives: From historical origin to present-day development. In K. Menzel, E. Lapshinova-Koltunski, & K. Kunz (Eds.), *New perspectives on cohesion and coherence. Implications for translation* (pp. 11–34). Language Science Press.

Schiffrin, D. (1987). *Discourse markers*. (Studies in Interactional Sociolinguistics 5) Cambridge University Press.

Schiffrin, D., Tannen, D. & Hamilton, H. E. (Eds.). (2001). *The Handbook of Discourse Analysis*. Blackwell Publishers.

Stubbe, M., & Holmes, J. (1995). You know, eh and other 'exasperating expressions': An analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language & Communication*, *15*(1), 63-88.

Takamura, R. (2020). Discourse marker well: A linguistic key to the well-being of human interaction. *早稲田大学高等学院研究年誌*, *64*, 79-103.

Tiedemans, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218.

Tsai, P. S., & Chu, W. H. (2017). The use of discourse markers among Mandarin Chinese teachers, and Chinese as a second language and Chinese as a foreign language learners. *Applied Linguistics*, *38*(5), 638-665.

Zenner, E., & Geeraerts, D. (2015). I'm queen of the world! (Semi-) fixed English expressions and constructions in Dutch. *Taal & Tongval*, *67*(2), 247-274.

Zenner, E., & Van De Mieroop, D. (2017). The social and pragmatic function of English in weak contact situations: Ingroup and outgroup marking in the Dutch reality TV show Expeditie Robinson. *Journal of Pragmatics*, *113*, 77-88.