Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Triennale Interclasse in
Lingue, Letterature e Mediazione culturale (LTLLM)
Classe LT-12

Tesina di Laurea

Formulaic language in TV series: an analysis of the use of lexical bundles in the American sitcom "How I Met Your Mother"

Relatrice
Prof.ssa Katherine Ackerley

Laureanda
Anna Cracco
n° matr.1233274 / LTLLM

Anno Accademico 2022 / 2023

Acknowledgments

Desidero ringraziare innanzitutto il relatore di questa tesi, la professoressa Katherine Ackerley, per la disponibilità, l'attenzione e la gentilezza dimostrate durante la stesura del lavoro.

Un grazie in particolare va alla mia famiglia. Ai miei sostenitori numero uno. Il loro supporto è stato essenziale, soprattutto nei momenti di sconforto. Grazie ai miei genitori, Liliana e Giuseppe, che mi sono sempre stati accanto e mi hanno appoggiato in ogni mia scelta. Grazie alle mie sorellone Federica Valeria ed Alessandra, che hanno sempre trovato una parola di conforto e di carica per spronarmi, riuscendo sempre a tranquillizzarmi nei momenti più difficili.

Un grazie di cuore ai miei amici. In particolare, alle mie meravigliose amiche, che con pazienza mi hanno ascoltato e hanno supportato ogni mio momento di gioia e di sconforto, cercando di strapparmi sempre un sorriso.

Ringrazio le mie compagne di corso, Giulia ed Anna, che sono diventate amiche speciali, che ci sono sempre state in questo percorso e con cui ho condiviso e, sono sicura, condividerò ancora tanto.

Infine, un ringraziamento speciale ai miei adorati nonni, anche ai miei due angeli, a cui dedico questo successo e che sono e saranno sempre una parte importante di me.

Voglio un bene infinito a tutti voi, grazie di essere nella mia vita.

Table of contents

Introduction

Formulaic language is a term for some expressions which are fixed in form, such as idioms, collocations, phrasal verbs, lexical bundles and many more. In particular, this dissertation sheds light on "lexical bundles" (Biber et al. 1999: 990), which are considered "a sequence, continuous or discontinuous, of words or other elements which is or appears to be fabricated" (Wray, 2002: 9). Moreover, lexical bundles, or recurrent word strings, are one of the key elements in increasing fluency of language and in second language learning. In most previous works, indeed, lexical bundles were identified and analysed, however little research paid attention to spoken discourse. In fact, the present paper fits into studies of these linguistic elements of formulaic language in spoken language. For this reason, I decided to focus my dissertation on this topic and to analyse the fixed expressions mentioned in a specific TV series to which I am emotionally attached "How I Met Your Mother". Overall, what I will do is analyse the frequency of lexical bundles in two specific corpora: Santa Barbara Corpus of Spoken American English, representative of authentic language, and the "HIMYM Corpus", representative of scripted language of fiction.

More precisely, this paper is divided into two main parts: the first provides the theoretical background of formulaic language, lexical bundles, and authentic spoken language in TV series (Chapters 1 and 2); the second part presents the practical and personal analysis of some of these fixed linguistic sequences found in the corpora chosen (Chapter 3). In detail, chapter one outlines the definition of formulaic languages and lexical bundles, describing their cultural and historical background as well as their functional and grammatical features and reporting the positive effects they have on fluency and mastery in language use. Subsequently, chapter two deals with the term "authenticity". In particular, it focuses on authentic spoken language in television, especially in TV series, and will reveal that spontaneous linguistic elements of real-life conversation are included also in scripted speech of fiction, which, as we will see, is a source for authenticity. After, indeed, an initial comprehensive definition and description of *authenticity,* this chapter observes the television discourse and the presence of linguistic elements of spoken language in the edited language of fiction in order to provide a sense of reality although it is a *non-authentic* and scripted language.

Once I have all the theoretical part of the dissertation, I will shift to the practical side, namely to the corpus-based study. First of all, the present work addresses the following research questions: “what are the most frequently used four-word lexical bundles in the TV series HIMYM?”, “what are the most frequently used four-word lexical bundles in the real-life conversations (Santa Barbara Corpus of Spoken American English)?” and “what are the similarities and differences in the four-word lexical bundles found in the two corpora in terms of Biber’s discourse functions (Biber et al. 1999)?”. To answer these questions some data from an existing spoken American English corpus (Santa Barbara of Spoken American English Corpus) are analysed, and a new corpus of American TV series conversation must be built, explored and compared with the first corpus. This corpus-study firstly sheds light on the frequency of three-to-five-word lexical bundles, and then focuses on some four-word lexical bundles of both the corpora analysed and their similarities and discourse functions. In the third chapter, indeed, I will first briefly illustrate “How I Met Your Mother”, then I will describe the corpora used, and the technological tools employed in the analysis. Initially, the frequency of three-to-five-word lexical bundles will be identified, using the AntConc program. After that, the third chapter investigates the frequency of the most frequent four-word lexical bundles (“I don’t know”, “I don’t think”, “don’t want to”, I’m going to”) between the two corpora; subsequently, their discourse functions, using Biber et al.'s (2004) functional taxonomy as a framework, will be identified. This part of the chapter will include examples for each corpus of the context in which lexical bundles occur and a subsequent explanation of the results found.

In essence, the main objective of this paper is to investigate the frequency of these specific linguistic expressions in both scripted language and face-to-face conversations in order to recognize the similarity between spoken American English and the language used in TV series, in this case “How I Met Your Mother”. This makes it clear that although television language is referred to as “non-authentic”, not spontaneous and planned, it is actually artificially designed to sound like real-life conversations and thus to include *authentic* spoken language, especially when it stages scenes of everyday life, as in the case of the series I have chosen to analyse.

Chapter one

Formulaic language and lexical bundles

This chapter intends to give a more detailed definition of the term formulaic language, by giving information concerning the background in which it developed and analysing the various features and functions. Subsequently this first chapter aims to deal with the phenomenon of *Lexical bundles*, a category within formulaic language, which is explained in depth in the last part of the chapter with various explanations and examples.

1.1 Formulaic language

Formulaic language is a term which refers to recurrent multi-word lexical items or "a system of prefabricated unit languages and phrases creatively knit together" (Győrfi, 2017:29). The linguistic units of formulaic language are called *formulaic sequences* and are describes as "multiple word phraseological units" (Schmitt, 2010: 117). Formulaic language has been referred to as "multiword sequences" (Butler, 2003: 179-208), "prefabricated patterns" (Hakuta, 1974: 287-297), "idioms" (Wray and Perkins, 2000: 3), "collocations" (Pawley and Syder, 1983: 208) or "lexical bundles" (Biber et al., 1999: 990). Despite the terminological variation, meaning that formulaic language "has been used with a with a multiplicity of meanings" (Myles and Cordier, 2017: 4), Wray tried to give a comprehensive definition of the term:

> A sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray, 2002:9).

Formulaic language is common in both written and spoken language, according to Wray (2002) indeed, there are chunks of language that appear repeatedly in daily life. Sinclair (1991: 110) developed the *idiom principle,* a crucial principle within the field of formulaic language which explains that, rather than being constructed piece by piece in small sections, language is constructed by putting together larger pieces of already formulated phrases.

Although not long ago there was little research being conducted on formulaic language, these days it is a language phenomenon that is quite well known among researchers and students in linguistics. The actual turning point was probably around 1970, when some structural linguists actually began to pay some attention to formulaic language. Up to that point the work on formulaic language had been conducted by people from a diverse range of areas of interest, from literary studies to anthropologists, to educational psychologists and neurologists, and again, to philosophers and sociologists. One of the first to focus on formulaic language was Parry (1928, 1930), when he turned attention to the "South Slavic tradition of public performance of epic poems". These lengthy poems were written and performed by people who were not literate, and Parry noted that formulas in such poems allowed the performance to be more fluent, rhythmic, and smooth. Considering that the early examinations of formulaic language focused on epic poetry and oral traditions in various cultures, anthropologists began to pay attention to this language phenomenon, and started analysing different types of spoken language from everyday speech to magical incantations to child language play and found evidence of formulaic language playing a strong role. Malinowski (1935) noted that the Trobriand islanders used fixed formulaic language in their magical incantations, while Hymes (1962) decided to apply the anthropological research methods to work in linguistics and, as a result, he investigated the performance routines and recurrent patterns in everyday speech. Furthermore, in the 1960s, the study of everyday communication grew to focus on the use of routine utterances to accomplish speech acts. Goffman (1971) began to analyse conversation as a discipline within linguistics and pointed out the fact that many conversational moves are accomplished by means of conventional word usage, an early type of formulaic language research. Neurologists have also paid attention to formulaic language, showing that damage to Broca's area of the left hemisphere reduced or eliminated the ability to use propositional speech but left the sufferers with the ability to use familiar expressions, formulaic sequences.

Additionally, it became evident to researchers that fluent speakers have more automatized chunks of language to use while speaking spontaneously. According to Chafe (1980), "speech is produced in bursts, in which vocalization is broken up by pauses at junctures of meaning and syntax" (Wood, 2015: 18). Moreover, as mentioned

above, the 1970s represent a turning point in formulaic language research, and various linguists pursued research in the area; in particular, lexicographers began to analyse multiword chunks, speech acts and pragmatics. Pawley and Syder published a landmark paper in 1983 where they discussed the fact that formulaic language has a role in language fluency and nativelike selection, and they also pointed out the tendency people have to use routine ways of expressing things despite the supposed infinite potential of language. Again, according to Pawley and Syder speakers acquire language in "chunks" (Pawley & Syder, 1983: 191), store and retrieve word strings often as wholes from long-term memory to fit the meanings and functions that arise in communication. Lastly, in 1991 Sinclair also developed the *open choice* principle, a somewhat similar idea, that "most texts are largely composed of multiword expressions that constitute single choices in the mental lexicon" (Sinclair, 1991: 110).

1.1.1 Characteristics of formulaic language

Schmitt (2004) suggests there are several characteristics of formulaic sequences established in literature. Despite the fact that formulaic sequences can include "an intervening variable slot" (Forsyth and Grabowski 2015: 7) which enables flexibility" (Wray, 2012: 246) of use, these sequences appear to be stored as "holistic units" (Siyanova-Chanturia, 2015: 285) and feature certain traits that allow them to be recognized.

To begin with, a major feature of this prefabricated language is its arbitrary nature, which means there is "no particular expression preferable among other systematic expressions to fulfil a specific communication function" (Schmitt, Crater, 2014: 1-22) Formulaic language, indeed, includes borrowed, common and understandable lexical items, notions, expressions that contribute to the formation of a "purely independent form of language" in order to fulfil a range of communication objectives (Schmitt, Crater; 2014: 1-22).

In addition to this, formulaic language, unlike the normal structure of natural language (also called novel communication), is seen as ready-made chunks and sequences of words", which are "stored as wholes in long-term memory" (Wood, 2002: 1). What makes formulaic language different from natural language are some defining features, such as the fact that these formulaic expressions are retrieved as "stereotyped

forms" (Van Lancker Sidtis, 2012: 64). First and foremost, formulaic expressions are stereotyped as they are an abstract creation and formulation by a community of persons with similar challenges. In fact, formulaic language is achieved through a combination of commonly known and well-acceptable cultural norms.

It also should be noted that formulaic language expresses a "conventional meaning", in the sense that formulaic expressions allude to ideas, practices and activities understandable by many (Van Lancker Sidtis, 2012: 64). In addition to this, research on the "pragmatics" of formulaic language shows that formulaic expressions are not independent of context; rather they are "sensitive to social conditions like discourse styles, formality index, social register and the structure of communication" (Ellis, Simpson-Vlach, Römer, O'Donnell, Wulff; 2015: 333-356). Formulaic expressions are, therefore, more narrowed with a specified application that suits social conditions, for example: greetings such as "Good Morning" or "Goodnight" are closely tied to contextual particulars as they are usually said at a specific time throughout the day. Overall, formulaic expressions apply to specific conditions and achieve defined communicative purposes.

Formulaic language features "a known status in that it is easily identifiable and well known". For this reason, even in case of brain/memory damage (aphasia or dementia), which affects the lower faculties, the brain is conditioned to store information in the long-term memory. This is due to the brain's ability to preserve a significant content for information conveyance. Therefore, formulaic sequences are frequently repeated. For this reason, some researchers created different frequency lists which focus on the most frequent formulaic sequences in various context. More precisely, Liu (2003) developed a list based on three different corpora of spoken American English (professional, media, and academic corpora) and mentioned some frequently used fixed forms such for example: "kind of" (meaning somewhat); "of course"; "make sure"; "deal with" or "as well" (Liu, 2003: 671-700).

1.1.2 The functions of formulaic language

According to some scholars, language is "largely formulaic in nature", and that "phraseological competence is an important part of nativelike, fluent, and idiomatic language use" (Paquot, Granger, 2012: 310). Formulaic sequences play some very

important roles in language; for this reason, they are widespread in both spoken and written discourse. One main point to bear in mind, indeed, is that formulaic language may convey various meanings and have several functions. Some of these include: stating a commonly believed truth or advice, defining discourse organization, maintaining conversations or even providing expressions which facilitate social interaction. However, in order to elaborate further on this issue, it is important to discuss individually the major functions of formulaic sequences.

According to Nattinger and DeCarrico (1992), formulaic sequences are "ubiquitous" in language use. In fact, Foster (2001) states that 32% of unplanned English native speaker speech consists of formulaic language, while Erman and Warren (2000) conduct a corpus study and conclude that 58% of language is formulaic. As far as pervasiveness is concerned, formulaic language is a feature of many languages, such as French, Swedish, Hebrew, Chinese and many more (Schmitt, 2010).

Using formulaic sequences is probably an important factor in the ability of proficient speakers to produce a fluent discourse and to decrease processing effort. Wood (2002: 31) claims that formulaic sequences appear to be "multi word-units, stored as wholes in long-term memory" [i.e., "lexical chunks" (Zhang, Lu, 2017: 74) "ready-made chunks" (Wood, 2002: 1)] and can therefore be processed and accessed more quickly and easily than the same sequences of words when generated creatively". In fact, these sequences of words are represented, processed, and accessed holistically and this process helps speakers to save mental effort so as to conceptualize and formulate the next stretch of discourse, simultaneously maintaining a certain pace and rhythm of speech.

As a matter of fact, from a cognitive perspective, the use of formulaic sequences in academic and everyday conversations allows the speaker to produce a discourse marked by a high language proficiency or fluency and to achieve communicative purposes (Schmitt, 2010; Wray, 1999). Formulaic language plays a fundamental role in communicative competence: "most speech acts such as expressions of apologies (e.g., *I'm sorry,* etc.), requests (e.g., *could you [please]* etc.), compliments, greetings, thanks and condolences tend to be achieved by using formulaic sequences" (Hatami, 2015: 116). Furthermore, formulaic sequences are commonly used in making discourse structure (e.g., *one the other hand, as a result*) and in conversation management (e.g.,

you know, I mean, you'll never believe this, but). In addition to this, these fixed and idiomatic sequences, shared within a community, can help learners achieve linguistic accuracy and avoid making errors; mastery and use of these sequences may reduce listener effort in conversation, thereby facilitating communication efficiency, efficacy, and allowing the speaker to be confident that what it is saying will be understood.

Finally, formulaic language also performs important social functions in various social contexts. The use of formulaic sequences reflects a strong identity and membership within a particular speech community. By using formulaic language new members will gain access easily into the community, and the use of formulaic sequences can enhance their communication with other members. In essence, formulaic language is crucial for accurate, appropriate, and overall fluent language use.

1.2 Identifying formulaic language

A more in-depth examination must be made with regard to categorization of formulaic language. As mentioned earlier, this prefabricated language includes various features; because of this, these prefabricated unit languages can be divided in different formulaic categories, including for example: idioms, phrasal verbs, collocations, and lexical bundles.

1. Idioms: a phrase or expression that, when taken as a whole, has a meaning that can't be deciphered by defining the individual words. Some examples are: "under the weather" (meaning "not feeling well") or "break a leg" (used to wish someone good luck).

2. Phrasal verbs: a verb (e.g., "it goes") becomes a phrasal verb with the addition of one preposition (e.g., "the light goes out", meaning "to turn off") or more (e.g., "she goes out with him" meaning "to date"), which completely change the meaning of the verb. Phrasal verbs describe a specific action in everyday speech and must be treated as distinct pieces of vocabulary that have to be learnt as a single unit of meaning: the verb "goes out" (meaning "to date") is as different from the verb "goes" as the verb "goes" is different from the verb "moves."

3. Collocations: they are defined as words that co-occur more frequently, and which mainly include word pairs of high frequency such as: adjective + noun

(e.g., fatal mistake) or verb + noun combinations (e.g., provide information). They are easily recognized and there is clear evidence that frequently occurring collocations are processed quickly.

4. Lexical bundles: they are "combinations of three or more words" which are identified "empirically in a corpus of natural language" (Cortes, 2006:392). Some examples of lexical bundles are expressions such as "I don't know", "take a look at", "on the other hand", or "as a result of", among many others.

Despite this clear and comprehensive distinction within such formulaic sequences, the actual identification of formulaic language in texts and discourses is fairly complex. Although some cases are clear such as idioms, phrasal verbs, and so on, many uncertain areas, such as *discontinuous* expressions, remained. Formulaic language, indeed, can also include discontinuous sequences of words, that Renouf and Sinclair called "collocational framework" and defined as a "discontinuous sequence of two words, positioned at one word remove from each other" (Renouf and Sinclair, 1991: 128). Similarly, Moon refers to these expressions as "lexicogrammatical frames" and states that "the realizations of one constituent vary relatively widely" (Moon, 1998: 145-6). Some examples of collocational frameworks are: "a * of", "for * of", "too * to", "beyond *", "many * of", "point * view" where the asterisk represents a variable slot in the frame that houses a "filler."

1.2.1 Criteria of FL

As a consequence of the discontinuous sequences mentioned above, Wray (2002) provided a set of criteria for determining if multiword combinations might be prefabricated. Formulaic sequences often begin with conjunctions, articles, pronouns, prepositions or discourse markers and, as a consequence, structure or form of the multi-word sequences is one such criterion (e.g., *as described previously, a wide range of* etc.). In particular, internal structure or compositionality of these word sequences is noteworthy and, as Wray observes, "the string is no longer obliged to be grammatically regular or semantically logical" (Wray, 2002: 33). Some formulaic sequences for example are not easily retrievable because of their lack of semantic transparency such as

the sentence "up the creek without a paddle" (which means "being in a difficult situation").

Another criterion which allows the identification of formulaic sequences is *fixedness*. This term refers to the tendency for prefabricated sequences to be of invariable form, although Wray does allow that "a large subset of formulaic sequences often have fillable slots" (Wray, 2002: 34). In fact, some formulaic sequences are considered fixed strings of words (e.g., *Ladies and Gentlemen*), while others are characterized by and have indeed fillable slots which allows them to be used flexibly (e.g., *it goes without saying that* etc.). These slots have semantic constraints, which means that they can only be composed of semantically appropriate words. The fixedness of formulaic sequences has been examined through recurrence of word sequences or frames, mostly through a corpus-based approach where computational algorithms are created to identify words that tend to co-occur across utterances, contexts, and interlocutors.

Lastly, other criteria relate to "phonological or prosodic aspects of the articulation of a sequence, including intonation contour and speed of articulation, and fluency criteria such as lack of internal pausing" (Wray, 2002: 35). Peters (1983), who was the first to use the term *phonological coherence*, states that these prefabricated utterances are "always produced fluently as a unit with an unbroken intonation contour and no hesitations for encoding" (Peters, 1983: 8). In addition to this, she includes what she calls *phonological coherence* as one of her lists of criteria for identifying formulaic sequences in child language. However, while there is empirical evidence to support the use of phonological coherence to identify formulaic sequences in child language, researchers must wait a few years in order to ensure that it is valid to directly apply the phonological criteria in the identification of formulaic sequences also in adult spoken language.

For a more in-depth explanation on the phonological coherence, it is important to mention the "theory of holistic storage", which assumes that formulaic sequences should form single intonation units, have less internal dysfluencies, such as hesitations and pauses, and require specific accentual patterns or focus distinction. With regard to the phonological aspect, indeed, the theory especially refers to formulaic sequences in child language: when a child pronounces an utterance, which features *phonological*

coherence, including "distintive intonation" (Peters, 1977: 560-73), an "unbroken intonation contour" (Peters, 1983: 8) and "arti-culatory fluency" (Plunkett, 1990: 1-17), it almost surely means that the whole sequence has been learned as a *chunk* and stored as *a holistic unit* in the child's mental lexicon.

1.2.2 Identification of formulaic language in spoken discourse

As far as spoken language is concerned, instead, Wray states that within spoken discourse, "it may simply be that identification cannot be based on a single criterion, but rather needs to draw on a suite of features" (Wray, 2002: 43). Phonological features such as speech rate, pausing and stress patterns are possible indicators of formulaic sequences in spoken language (Read & Nation, 2004). More relevant researches have been conducted in the field of spoken language by focusing on the links between fluent speech production and formulaic language. To refer back to one of the main roles of formulaic sequences, Wood has analysed formulaic language in speech with second language learners, finding that increased "use of formulaic language improves speech fluency", and that they appear to use this language "to facilitate speech fluency by relying on one sequence, repeating a particular sequence, stringing together multiple sequences, and using them as self-talk or rhetorical structuring devices" (Wood, 2009: 39-57). However, one of the earliest observations about the nature of spoken discourse was that of Pawley and Syder (1983: 191-225), who described the way clauses tend to be chained. They concluded that everyday fluent conversational speech is composed largely of strings of more or less independent clauses, without much grammatical integration. For example, subordination is only present in limited amounts in spontaneous speech (Pawley & Syder, 1983: 202-204). According to Pawley and Syder, indeed, clause-chaining is most effective in narrative speech:

> With the chaining style, a speaker can maintain grammatical and semantic continuity because his clauses can be planned more or less independently, and each major semantic unit, being only a single clause, can be encoded and uttered without internal breaks […] may speak, then, of "a one clause at a time facility" as an essential constituent of communicative competence in English: the speaker must be able regularly to encode whole clauses in their full lexical detail, in a single encoding operation and so avoid the need for mid-clause hesitations (Pawley & Syder, 1983: 203-204).

In conclusion, a clue about one prime function of formulaic language in spoken communication is related to the fact that by chaining clauses in conversational and narrative speech in English a speaker should be able to encode whole clauses and avoid hesitations in mid-clause. Much of everyday speech is formulaic, and Pawley and Syder (1983: 205) suggest that memorized chunks form a high proportion of the speech of everyday conversation may bring many benefits: "if speech is formulated and articulated word-for-word, a speaker's attention is freed to focus on rhythm, variety, combining memorized chunks, or producing creative connections of lexical strings and single words".

1.3 Lexical bundles

Formulaic language and *formulaic sequence*, appear to be proper terms related to "formulaicity" of language (Wray, 2002: chap. 17). As mentioned earlier, there are many categories of formulaic language, including collocations, idioms, phrasal verbs and lexical bundles. Many researchers turned attention to the relevant category of "lexical bundles", as one approach of analysing language formulaicity. Biber et al. (1999) introduces for the first time the term "lexical bundles" in the "Longman Grammar of Spoken and Written English" and refers to it as "sequences of word forms that commonly go together in natural discourse" (Biber et al., 1999: 990). Similarly, according to Hyland (2008), lexical bundles are "word strings that happen more frequently than expected by accident in the text" and not only "help to shape the meaning" but also "help the readers to distinguish across different genres" (Rezaie, Farahani, Masoomzadeh, 2020: 130). A more detailed insight to the lexical bundles is given by Altenberg who highlights the fact that these bundles are formulaic expressions with three common features: "semantic transparency, fragmental grammatical structures, and pragmatic specialization" (Altenberg, 1998: 101-122). He investigated some word combinations in the "London-Lund Corpus of Spoken English" in order to determine whether there is a certain amount of formulaic language in the spoken corpus, for example: "I do not think", "do you know", "on the other hand", "yes I did" were identified as recurrent word clusters. Moreover, results showed that most of the lexical chunks are transparent in meaning although the majority of word combinations have

incomplete grammatical structures, with 76% being clause constituents and 14% incomplete phrases.

Although there is no indication in literature that lexical bundles are stored or retrieved as wholes, according to Biber et al. (1999: 991-993) "lexical bundles extend across structural units". Biber, in fact, recognizes various bundles structural types, which are sensitive to different genres or registers. Results show that 44% of lexical bundles in conversations take the form of verbal and clausal units, as in *"*I don't know why" and "I thought that was". On the other hand, lexical bundles in academic writing are more likely to be nominal, usually in the form of noun phrases and prepositional phrases such as "the size of the". As far as grammatical features are concerned, despite variations in grammatical structures of lexical bundles, they share the incompleteness in their grammatical structures.

1.3.1 The functions of LBs

Returning back to the pioneer of lexical bundles Biber (2004), who refers to lexical bundles as "the most frequent, recurrent, multiword sequences in a register, which can be identified strictly on the basis of frequency" (Biber, 2004: 399), decided, along with Conrad and Cortes (2004), to describe the use of lexical bundles in two university instructional registers: classroom teaching and textbooks. As a result, they developed a functional categorization of the use of bundles. Finally, in 1999 the first structural classification was proposed depending on both the structures and functions of the lexical bundles. They identified three main groups:

1. *Stance expressions* are used to "express attitudes or assessments of certainty that frames some other proposition" (Biber et al., 2004: 384). These bundles split into two major semantic categories:
    - Attitudinal/modality stance bundles
    - Epistemic stance bundles.

The former reflects the "speaker attitude toward action or event described in the following proposition" (Biber et al. 2004: 390) and it is divided, itself, in four functional sub-categories: desire (i.e., *I don't want to*); obligation (i.e., *you need to know)*; intention/prediction (i.e., *we are going to*); ability (i.e., *going to be able)* bundles. On

the other hand, epistemic stances deal with expressions of certainty or uncertainty (i.e. *the fact that the, let's assume that)*.

2. *Discourse organizing bundles* "reflect relationships between prior and coming discourse" (Biber et al., 2004: 384), and are categorized as:
    - *topic introduction and focus bundles*, which are employed to introduce a new topic, frame a section of a text or to emphasize a main point. (i.e., *take a look at, want to talk about).*
    - topic *clarification/elaboration bundles,* which are used to further explain and elaborate topics. (i.e., *as well as the*, *and on the other hand*, i*t turns out, what happens is, in order to* and *as a result).*
    - *discourse markers*, which include both "connectives" (*as well as, in other words*) and "interactive devices" and "formulas" (*thank you very much*) (Simpson-Vlach & Ellis, 2010).

3. *Referential bundles*, the largest of the groups, are employed to "make direct reference to physical or abstract entities, or to the textual context itself, either to identify the entity or to single out some particular attribute of the entity as especially important" (Biber et al., 2004: 384). These bundles are divided into five groups depending on the pragmatic functions they have:
    - *identification and focus referential bundles* which include typically expository phrases (i.e., *such as the, referred to as* and *an example of)*. These bundles also focus on the readers'/listeners' attention on the noun phrase after the bundle (i.e., *for those of you who*).
    - *attribute specifying bundles*, which can include:
      -quantity specifications in order to specify the quantity of the following noun phrase (i.e., *a lot of the, equal to*).
      -framing devices in order to frame the tangible (i.e. *the amount of, the size of, the level of)* or intangible (i.e*.,* the nature of the, *based on the, in such a way)* properties of the noun phrases.

- imprecision bundles which suggest vagueness of reference in conversation as well as in academic writing and include a small number of prefabs (i.e., *things like that, and so on)*.
- time/place/text referential bundles which refer to physical locations or "spatial reference points in the discourse" (Simpson-Vlach & Ellis, 2010).

As far as the three functional categories described above, overall, the distribution of lexical bundles across functional categories in Biber's studies show that "referential bundles" (Biber, 2004: 390-393) are the most common type of function for lexical bundles in the textbooks, whereas "discourse organizers" (Biber, 2004: 390-393) were the least common. Moreover, within the category of referential functions, it appears that intangible framing and quantity specifications represent the largest categories.

1.3.2 Lexical bundles across registers and corpora

Comparative studies of lexical bundles are largely carried out along two main dimensions: the use of lexical bundles across disciplines and across registers. "Discipline" is an important variable influencing the use of lexical bundles (Cortes, 2004: 397-423). Biber (2006), in a corpus-based analysis of university language concerning the study of lexical bundles in textbooks, discovered the use of these prefabs in academic disciplines differ: for example, natural and social sciences use lexical bundles more than the humanities. In addition to this, according to Cortes "social science texts" make use of a large number of bundles with an "embedded *of* phrase" in order to identify the "logical relations" in the discourse (Cortes, 2004: 397-423). To continue, Biber (1999) affirmed that lexical bundles are "theoretically attached to assigned registers" and for this reason decided to analyse the use and frequency of lexical bundles in non-identical registers (i.e., conversation and academic prose), and he consequently observed that register variation was one of the crucial variables influencing the use of lexical bundles. Moreover, Biber concluded that, in terms of structure, "bundles are more clausal in conversation but more phrasal in academic prose" (Biber, Johansson, Leech, Conrad, Finegan, 1999), whereas the use of lexical bundles in classroom teaching reflects a combination of characteristics typical of both

spoken (through the use of stance and discourse organizing bundles) and written registers (by using referential bundles). Subsequently, Biber, Conrad, and Cortes (2004) decided to conduct a study, that compared the use of "multi-word sequences in two important university registers: classroom teaching and textbooks" (Biber, Conrad, Cortes, 2004: 371) to earlier research on bundles in conversation and academic prose just mentioned. The results for classroom teaching revealed that many more bundles are employed in this context rather than conversation, academic writing, or textbooks; and the use of stance bundles in teaching far exceeds the one in conversation, as well as the use of referential bundles in academic writing. In conclusion, this research indicates that lexical bundles are not an accidental linguistic choice and are ubiquitous in both spoken and written texts, where they serve basic discourse functions related to the referential framing, discourse organization and expression of stance.

Overall, in essence, lexical bundles occur across a range of texts in several disciplines but also in a corpus. This prefabricated language combinations of two or more words are identified in a corpus of natural language by means of corpus analysis software programs. In fact, it is important to highlight that lexical bundles are researched using particular methods which rely on pre-selected lengths and frequency rates and in focus exclusively on function and frequency. The "lexical bundle approach", in fact, is such an important analysis of formulaic sequences created by Biber and collegues' that relied on "frequency" (lexical bundles should occur more than 20 times in a million words to qualify) as the "identifying criteria" but also on the continuous nature, and length of what they called "lexical bundles" (Biber & Conrad, 1999: 181-189). This research shows that frequent uninterrupted expressions and lexical bundles longer than two words tend to be more employed (frequency) for discoursal functions, thus more essential for discourse (Conrad & Biber, 2004). Additionally, another identifying criteria of lexical bundles is the focus on the different lengths of these multi-word sequences, as mentioned before. For instance, Biber's (1999) quantitative analysis of the 40 million-word Longman Spoken and Written English Corpus identified that "about 28% of the words" occurred in "3- and 4- word lexical bundles" (Biber, 1999: 993-4). In conclusion, considering that "four-word lexical bundles will show longer lexical bundles" (Biber, Conrad, Cortes, 2004: 371-405) (e.g., *if you know what, you know what I, etc.)*, it is evident that most of lexical bundle

research is concentrating on four-word lexical bundles, as in Cortes (2004), Biber and Barbieri (2007), and Ädel and Erman (2012) to mention a few.

In conclusion, this first chapter illustrated how frequently occurring fixed expressions such as formulaic language, and especially, lexical bundles are considered an important indicator of successful language users, in both written and spoken discourse. In fact, knowledge and proper use of these frequently occurring multiword sequences ensures fluency and effectiveness in language use, or even native-like language competence. The following chapter, instead, intends to focus on spoken language and, in particular, to authentic spoken language on television, trying to expound on its peculiarities and benefits as well as linguistic challenges during language learning. In essence, in this second chapter the literature review on the analysis of authentic language in TV series has a central role and will also have the function of giving a general outline of what I will expound in chapter three.

Chapter two

Authentic spoken language and TV series

This second chapter aims to focus on authentic spoken language in television and, especially, in TV series, trying to expound on the peculiarities of this language as well as benefits and linguistic challenges during language learning.

2.1 Authenticity

Authenticity is a term with no clear definition. Porter and Roberts refer to authentic texts as those "instances of spoken language which were not initiated for the purpose of teaching […] not intended for nonnative learners" (Porter and Roberts, 1987: 177), while Rogers and Medley observe authentic texts are those that "reflect a naturalness of form and an appropriateness of cultural and situational context" (Rogers and Medley, 1988: 468). Despite this, the adjective *authentic* refers to the idea of language realism, which reflects the reality. One clear example of authentic language is conversation, which takes place within a social context and is a spontaneous discourse depending on that context. In conclusion, authentic takes on the meaning of language use in real life, as opposed to 'non-authentic', which describes scripted dialogue of television. Despite this, we will see that linguistic features of real-life conversations are included in the language of television, which aims to represent reality and is therefore increasingly a source of *authenticity*.

2.1.1 Identifying and classifying authentic language

There are different criteria to identify authentic spoken materials. To achieve an ideal authentic speech, it should contain features equated with those of natural discourse, such as:

- Natural speed and natural phonological phenomena (such as natural pauses and intonation, the use of reduction, assimilation, elision, etc.)
- High-frequency vocabulary, colloquialisms (such as short formulaic utterances, running jargon, etc.)
- Hesitations (pauses, paralinguistic markers such as (nervous) laughter or coughing)

- False starts (an utterance is aborted and then restarted with a new train of thought)
- Self-corrections
- Minimal responses (forms such as “mmhmm”, “yeah”, “uh-huh”, and “right”) which are uttered by a listener during a speech and signal a certain level of engagement with the speaker. (Rost, 2002: 125).

If the spoken language includes most of these features, then it is more natural and hence more authentic. In addition to this, with regard to the classification of authentic (close to natural) discourse is concerned, multiple scholars proposed different classifications which reflect their personal view. Rings (1986) for example, states that native speakers’ spontaneous conversations are the most authentic, while conversations composed for and “printed in textbooks” are considered “inauthentic” (Al-Surmi, 2012: 675). Overall, then, the classifications of the authentic spoken materials are mainly based on scholars’ intuition and personal judgment.

2.2 Authentic (spoken) language in TV

Taking a closer look into television language, the right term to represent this kind of non-authentic language, which contains features of the authentic one (spontaneous conversations), is *scripted language*. This term refers to a written text which is ‘written-to-be-spoken’. In the context of television, it would refer to the actual script for a film or TV program, so it refers to speech which has its origins in writing, in the sense that it is based on a written document created by scriptwriters. Scripted language is the concrete spoken realization of the written text, i.e. the spoken language as it is realized by the speaker, such as the presenter, the newsreader, the actor, etc. This spoken text would be transcribed to be analysable (transcripts). Additionally, scripted dialogue, namely a written or spoken conversational exchange between two or more people, has been studied as a representation of face-to-face conversation. Rey (2001), for example, used the American television series “Star Trek” for a study of language (2001: 138), stating that “the language used in television is obviously not the same as the unwritten language, but it represents the language that scriptwriters imagine real women and men produce”. Similarly, Biber and Burges (2001: 158) describe the artificial dialogues of

fiction and theatre as "useful representations of historical spoken language", while, on the other hand, Tannen and Lakoff (1994: 139) argue that "artificial language can represent an internalized model or pattern for the production of conversation". Given the limited availability of corpora of naturally occurring conversation, some scholars (e.g., Washburn, 2001) have suggested using television language (particularly situation comedies) for the purpose of teaching and learning English, as the best way to study this language and describe the characteristics of conversation is through the analysis of naturally spoken data.

For this reason, one of the many resources for authentic materials, particularly for spoken English, is Television. Although, indeed, language use in the entertainment system, particularly in fictional television and cinema, is considered 'non-authentic language', as it is deliberately non-spontaneous, television language best captures the core linguistic features of natural conversation, due to one important aspect that needs to be given attention when talking about language of television: "code of realism". This term according to Bubel (2006), refers to the "imitation of reality, which holds not only for language but for all elements of the film text, is typically achieved with the help of specific filmic conventions" (2006: 43). The author refers to the fact that there are common conventions that must be followed in terms of staged dialogue, but most important that realistic dialogue is a relevant part of modern television language. In fact, when typical elements of spontaneous conversation (e.g., *self-correction, hesitations, or minimal responses* etc.*)* appear in scripted dialogue, they consequently make the script more natural and less simulated. From this follows the fact that linguistic elements such as *lexical bundles* (analysed in the next chapter) used in spoken *authentic* language will be easily comparable and very similar to those found in television language, as this is a representation of spoken discourse.

For this reason, spoken TV materials (such as soap operas and sitcoms) are widely relevant and interesting for the language learner. As regards authenticity, it has been suggested that using relevant authentic materials, such as TV discourse, may offer learners the chance to encounter realistic (spoken) language beyond the confines of the foreign language classroom, which is the type of language they are most likely to encounter outside of an institutionalized context (Grant & Starks, 2001). Consequently, spoken TV materials are suggested for listening fluency development, accessing to

native speaker models, listening comprehension, speaking fluency and vocabulary learning.

2.2.1 Linguistic challenges of TV discourse

As already stated in the previous paragraph, TV discourse (TVD) is considered a rich and authentic source that resembles naturally occurring informal conversation, and a medium organized around the rhythms of speech, which includes spoken discourse. As far as television dialogue is concerned, there are distinctive characteristics of this spoken language compared to natural speech, which are imposed by the televised medium. As Cameron suggests "broadcast discourse has special characteristics which arise from the nature of the medium and the relationship it produces between speakers and addresses" (Cameron, 2001: 26). For example, in TV discourse there are almost "no overlaps", to avoid the possibility of misunderstandings by the audience, and at a "discourse level", there are "fewer repetitions and interruptions" than in natural conversation (Quaglio, Biber, 2006: 717). In fact, the most important feature in film and television language is fictionality. The fundamental linguistic system is the same, whether language is fictional or non-fictional (Baumgarten, 2005: 85), what is crucial, however, is what the fictionality implies. Fictionality is part of the discourse of television language and usually involves prepared (i.e., scripted) dialogue, which is intended to entertain the audience. In conclusion, despite the differences mentioned, the general similarities between television dialogue and face-to-face conversation suggest that television has the potential to provide researchers and teachers with a convenient source of spoken language data. Moreover, TV discourse (TVD) is considered close to naturally occurring conversation, from both linguistic and language-educational perspectives. Dose indeed, has called TVD an "auspicious compromise between artificial textbook dialogues and the overwhelming 'messiness' of genuine language data" (Dose, 2012: 103), as TVD lacks performance errors and contains fewer instances of overlapping speech, which may be hard to process by language learners, especially in earlier stages of proficiency. Furthermore, the informal nature of TV discourse (TVD) contributes to a sense of realism in television language, making this language use preferable even to textbook dialogues in terms of naturalness.

Referring to natural discourse Quaglio (2009), for example, conducted a study describing the language of the American situation comedy *Friends* and comparing it to this kind of speech. The data for this corpus-based investigation were provided by the transcripts of the television show (all the nine seasons) and the American English conversation portion of the Longman Grammar Corpus. The study of the situation comedy *Friends* reveals that the language in *Friends* shares the core linguistic features typical of natural conversation. In fact, the analysis shows that Friends shares many of the typical linguistic features of face-to-face conversation, but at the same time, includes some distinctive characteristics of fictional language, such as those mentioned in the previous paragraph (no overlaps, repetitions etc.). While he found that the language in the sitcom is less vague and more emotional and informal than naturally occurring speech, Quaglio states that "the use of television dialogue as a surrogate for natural conversation for the analysis of certain linguistic features seems perfectly appropriate" (Quaglio, 2009: 149). In conclusion, the overall impression remains that this TV sitcom represents spoken English and especially that TV language is strikingly similar to that of natural conversation, which includes not only elements of spontaneous conversation, but can also reveal a similarity in the choice of the type and discursive function of linguistic elements such as the *lexical bundles* mentioned in the first chapter.

In addition to this, over the years, some quantitative corpus-based analyses have shown that certain linguistic features are given much more importance rather than other features, when TVD is compared to unscripted conversational data. Based on Mittmann (2006) investigation of single- and multi-word items in a 50,000-word corpus of television series as compared to natural spoken English (Longman Spoken American Corpus), she notes a variety of differences. For instance, she discovers that greeting expressions (e.g., *hi, I'll see you*), interactive markers (e.g., *please, thank you*), and more markers of expressiveness (e.g., *look* and *listen*) are "overrepresented" in the TV series (Mittmann 2006: 577-578), while discourse markers are underrepresented in her television corpus. In conclusion, the author would like to convey a pessimistic view of film language, as a substitute for natural language in linguistic study and language teaching:

> The language of films cannot legitimately be studied to find out about the language of everyday dialogue. Certain single words and prefabricated expressions are overrepresented

> in the films, while others occur far less frequently. While it is unlikely that this will distort linguists' views of such interactions, it nonetheless means that language learners who attempt to improve their English through watching films will meet with a somewhat different kind of speech in the country itself (Mittmann 2006: 578).

Television has a limit in making a particular language clearly understood, even if English language learners (ELL) understand all the words in a conversation. Consequently, it is important considering potential difficulties that learners may have when confronted with authentic (non-scripted) conversational material. Spoken language in TV may contain varieties of English to which learners may not be accustomed and, consequently, the language that these mediums contain often is not easily accessible. Some linguistic obstacles such as "the rate of speech" (Wang, Yin, Zeng, 2015: 1) as well as "unfamiliar speech" (Mitterer & McQueen, 2009: 3) or the use of "reduced speech" (Meskill, 1996: 190), increase the difficulty of understanding the language itself. The most significant linguistic challenge facing scholars is figurative language, which uses words and expressions which have meanings that are different from their literal interpretation, such as: idioms, metaphors, irony, similes, and symbolism. This kind of language is frequently used in American television series (AMTVS), and it is more challenging to teach systematically to English learners because "figurative language typically arises from the role of phrases in discourse rather than from the sum of their grammatical and lexical parts" (Charteris-Black, 2002: 104). Even though figurative language often impedes learners' ability to understand those aspects of the dialogue, according to the research, AMTVS seem to be particularly relevant in teaching and a "multimodal English learning medium" (Kimball, 2018: 13) through which learners can increase their listening comprehension, vocabulary acquisition and acquire informal language.

2.2.2. Benefits of watching television series

Despite TV discourse in language education having been widely recognized as a tool to improve listening comprehension and vocabulary development, the area of grammar has remained less investigated. This may be since TV discourse has stereotypically been associated with "ungrammatical" content in terms of highly informal and non-standard usage (Werner, 2020: 4). As far as informality is concerned, in style, it is mostly

regarded as incorrect use of language, sometimes perceived as 'bad' language or 'slang' also because it is often spoken with a regional accent and dialect. This is surprising, as the suitability of using TV discourse (TVD) for "teaching pupils or students about many different aspects of language use, including expressions that might traditionally be neglected in textbooks or classroom teaching" (Bednarek, 2018: 245), that is, language variation, which is includes informal spoken and non-standard features. Over the years, indeed, TV discourse is well suited to illustrate varieties of English, in specific registers, with a particular view on informal spoken usage.

Moreover, one important issue to focus on, is the large number of informal language features that may be acquired through TVD (e.g., sitcoms, AMTVS etc.), such as idiomatic expressions, phrasal verbs, reductions and "phrasal chunks" (Hilliard, 2014:4) or "fixed lexico-grammatical units" (Werner, 2020: 9). As mentioned in the first chapter, chunks are fixed words or phrases that can combine with other elements but act as ready-made lexical units of language. In addition to these, informal lexico-grammatical features have been identified in TVD to "artfully and selectively simulates naturalistic speech, creating realism" (Bednarek, 2018: 180). They are considered as non-modifiable lexico-grammatical items, which are characteristic of informal and are included in the television language. These elements of spontaneous conversation are analysed in research conducted by Cullen & Kuo (2007), who decide to list the different functions for these phrasal chunks.

1. To create vagueness: Cullen and Kuo (2007) mention the particles "sort of" *and "*kind of*"* or "stuff like that", as conversational hedging devices that serve to create vague reference. They are items "used frequently at the end of utterances in conversation to allow the listener to identify a general set of items based on the characteristics of the items given before the tag" (Cullen & Kuo, 2007: 370). As mentioned above, they are used to hedge statements, as shown in these examples:

   -You meet a lot of people at Comicon, fan expos, *and things like that*. (Lost Girl, Flesh and Blood, 2012)

   -Listen, you got to tell me when you're grounded *and stuff like that* (Nashville, Your Wild Life's Gonna Get You Down, 2014).

2. To modify and show politeness: Yet another feature with a hedging function is what Cullen and Kuo refer to as the "modifying expression" (Cullen & Kuo, 2007: 370) *a (little) bit*, which serves to politely qualify adjective and noun phrases. Relevant examples are widespread in the TVD data, such as:
   - So, I think I'm *a little bit* justified. (I Hate My Teenage Daughter, Teenage Party, 2013)
   - Well, I got off to *a bit* of a rocky start with Cliff. (The Deep End, Pilot, 2010)

3. To mark discourse structures: A last feature set to consider are the discourse markers *"*you know" *and "*I mean"*,* referred to as inserts. These are regular features of interactive conversation, which are usually outside the syntactic structures they refer to. If "you know" serves the function of a shared knowledge or a pause marker to increase the speaker's discourse planning time, "I mean", on the other hand, is most often used when a speaker clarifies or rephrases, as these examples illustrate:
   - A: Until we get married a third time, you guys will never have to see each other again.
   B: Well, *you know*, actually that's not the case (The Big Bang Theory, The Conjugal Conjecture, 2016).
   - A: You don't like the idea?
   B: No, *I mean*, yeah, yeah, uh… absolutely. (Damages, Your Secrets are Safe, 2010)

Overall, Cullen and Kuo list with related examples are relevant for the purposes of this dissertation as it reveals how spontaneous elements of conversation will be used also in scripted speech of television language. Phrasal chunks, more specifically *lexical bundles*, will be the subject of the discourse analysis presented in Chapter three.

A matter to be further investigated, concerning spoken language on television, are American television series (AMTVS), considered as a popular medium of studying English for English language learners (ELLs) due to their authenticity. According to Kimball (2018) the use of AMTVS, that are more accessible now with the advent of Netflix, and YouTube and other digital platforms, can benefit several fields: from personal vocabulary, "informal, figurative, and formulaic language acquisition" as well

as their "listening and reading comprehension" (Kimball, 2018: 1). AMTVS is a "multipurpose" digital tool, which encourages "autonomous learning" and speaking in general (Kimball, 2018: 8). Over the years, various scholars have decided to analyse the relationship between TV series and language study, to understand the usefulness of this set of programs in language learning. Y. Wang (2012), for example, used AMTVS, to intentionally teach vocabulary to twenty-eight ELLs in Taiwan, by choosing three sitcoms (*How I Met Your Mother, Reba* and *King of Queens*) as the dialogue spoken in the series refer to the characters' "daily lives" (Y. Wang, 2012: 221). Wang (2012) surveyed students and conducted individual interviews to understand their perspectives and thoughts on watching the American sitcoms. Results showed that the majority of students agreed that using AMTVS helped them to learn vocabulary (e.g., words and idioms) that they stated to have never encountered in their textbooks where the misrepresentation of spoken language and the conversational grammar lead to learners' speech sounding unnatural and too formal. Additionally, Sockett's (2011) conducted research and noted that viewing, that learners engage during their spare time, influences their formal learning environments and aids second language acquisition. For this reason, Sockett (2011) based on the four most frequently viewed TV series online *(Dr. House, How I Met Your Mother, One Tree Hill* and *Lost*) by English learners interviewed, searched the transcripts of these series for the most frequently used four-cluster word phrases. Subsequently, he compared his results with the cluster word frequency word list on the British National Corpus (BNC) and determined the frequently occurring clusters to be "relevant examples of target language" (Sockett, 2011: 12) which reflect the language of real-life conversations, such as reduced speech, intonation, stress patterns, hesitations, and pauses.

In conclusion, this chapter has given basic notions of authentic language, defined their characteristics and also compared it to the scripted speech of television language. In particular, the focus shifted to the possibility of finding authenticity and spontaneity of spoken language in the TV series, and especially of finding a great similarity between linguistic elements, such as *lexical bundles,* of real-life conversations and of scripted language. In essence, in this chapter the analysis of authentic language in AMTVS has a central role and will also have the function of giving a general outline of what will be discussed in chapter three. The following and last chapter, in fact, intends to provide a

personal corpus-based analysis on the use of *lexical bundles* on TV series “How I Met Your Mother”.

Chapter three

A corpus-based study

As mentioned in chapter two, language used in real life is defined *authentic,* whereas *non-authentic* language consists of productions of conversation found in the entertainment system and refers to scripted speech (e.g., scripts of movies and TV series). In spite of this, the language used in scripts needs to be much more natural to represent reality and real-life conversations and spontaneity. In order to substantiate this, I chose to analyse the famous American sitcom "How I Met Your Mother", created by Craig Thomas and Carter Bays. The American TV series aired on CBS 2005 to 2014, for a total of nine seasons. The series follows the main character, Ted Mosby, in 2030, who tells his son and daughter the events that led him to meet their mother by telling the many adventures with his group of five friends (Ted, Marshall, Robin, Barney and Lily), who live and work in the New York city of the early millennium.

In the next pages I will first introduce the methodology involved in the analysis, describing the corpora and the analysis tools used, and the specific elements analysed. Through corpus analysis I will compare the frequency of lexical bundles in the first three seasons of HIMYM to the one in the "Santa Barbara Corpus of Spoken American English" (SBCSAE). Finally, this corpus analysis will focus in more detail on four most common lexical bundles in spoken American English, obtained from the analysis of two corpora, which will also be analysed and compared (using examples) according to their discourse functions (Biber et al. 1999). In the end I will answer the following questions:

1. What are the most frequently used four-word lexical bundles in the TV series HIMYM?
2. What are the most frequently used four-word lexical bundles in the real-life conversations (Santa Barbara Corpus of Spoken American English)?
3. What are the similarities and differences in the four-word lexical bundles found in the two corpora in terms of Biber's discourse functions (Biber et al. 1999)?

3.1 Methodology

In the present study, I will identify and analyse linguistic sequences that have a variety of labels, such as “lexical bundles” (Biber & Conrad, 1999), “clusters” (Scott, 1997) or “N-grams” (Stubbs, 2003). I will initially focus on three to five-word sequences and subsequently I will analyse the most common four-word linguistic sequences as they are more frequently used in spoken English. More precisely, this chapter will give space to a corpus-based study, i.e the investigation of corpora, i.e. collections of (pieces of) texts that have been gathered according to specific criteria and are generally analysed automatically. For this reason, in this chapter, two corpora will be compared to look for typical lexical bundles of speech. The corpora chosen for this corpus analysis include the “HIMYM corpus” and the “Santa Barbara Corpus of Spoken American English”, whose texts are examined using an analysis tool called AntConc. In fact, the purpose of the analysis with this program is precisely to observe the frequency of these linguistic elements of spoken American English in the authentic languages of daily conversations and the American TV series HIMYM.

3.1.1 The Corpora

For this study two corpora are used: the “Santa Barbara Corpus of Spoken American English” and the “How I Met Your Mother Corpus”. The first corpus employed is the “Santa Barbara Corpus of Spoken American English”, a corpus of spoken American English developed by researchers in the Linguistic Department of the University of California, Santa Barbara in 2000. The Santa Barbara Corpus of Spoken American English (SBCSAE) is a corpus of approximately 249,000 words, including hundreds of recordings accompanied by a transcript of naturally occurring spoken interactions from all over the United States (https://www.linguistics.ucsb.edu/research/santa-barbara-corpusi). The most represented form of interaction is face-to-face conversations between friends (gossip) or family members, arguments, on-the job talk and more. This corpus will be used in the investigation of lexical bundles in authentic language as it is a valid representation of spoken English, consisting of natural, spontaneous conversations, recorded in a systematic way. Consequently, the SBCSAE turns to be relevant for discovering how people really use English, not how they are supposed to use it or how they use it when we are writing. Subsequently, as a representation of

television language, exclusively for the present study, a new corpus was created with the unofficial scripts easily accessible from the internet for the TV series How I Met Your Mother. This new corpus includes dialogues and conversations between the characters from the first three seasons of the series (64 episodes). As mentioned above, the spoken interactions found in this new corpus are not spontaneous, as they are part of a scripted language, mentioned in the chapter two. Despite this, the presence of linguistic elements of spoken English (e.g., lexical bundles) make a comparison with the "Santa Barbara Corpus of Spoken American English" possible, also considering that they are both in American English.

3.1.2 Technological tools: Sketch Engine and AntConc

In the first place, the technological platform that allowed me to conduct the discourse analysis is Sketch Engine, developed by Lexical Computing CZ s.r.o (2003), which was useful for the creation of my personal corpus of the TV series How I Met Your Mother. Firstly, I collected the transcripts of the first three seasons of the American sitcom from a website (https://subslikescript.com/series/How_I_Met_Your_Mother-460649). After that, I copied those transcriptions in a Word document and deleted unnecessary data such as author names, figures or page numbers, not to confound the results of the analysis. Lastly, I used the digital platform Sketch Engine, to create a new corpus which includes the files I gathered. In fact, after compiling the corpus in Sketch Engine, the files were converted to the text files for the analysis and the "HIMYM corpus" of about 190,000 words was created.

AntConc is a freeware and multi-purpose corpus analysis toolkit, developed by Laurence Anthony (2014), for carrying out corpus linguistic research such as texts analysis, concordance and, what is most interesting for this dissertation, cluster and lexical bundles analysis. AntConc was used in this research to discover lexical bundles described in the first chapter as "combinations of three or more words" identified in a "corpus of natural language" (Cortes, 2006: 392). AntConc provides a tool to investigate multi-word units using section "Cluster/N-Grams". The terms N-grams and clusters are used especially as a reference word for multi-words expressions in the software for corpus analysis, i.e., AntConc. In the context of corpora, N-grams typically refer to tokens, namely the number of individual words occurring in the text. The N-

gram tool in AntConc produces frequency lists of sequences of tokens (words) and orders them alphabetically. The search term can be determined with cluster size as minimum and maximum words. It is also possible to specify a minimum frequency threshold for the clusters generated, the multiword sequences analysed at the end of this chapter are *the four-word strings* occurred minimum four times found in conversational American English (Santa Barbara Corpus) and in the TV series HIMYM. Expressions such as "I don't know" or "that's what I" can be considered lexical bundles, they do not have equal structures and are used in different contexts according to their discourse function (e.g., stance expressions, discourse organizers, referential bundles).

Figure 1. A screenshot from AntConc for Lexical Bundles search of the HIMYM Corpus

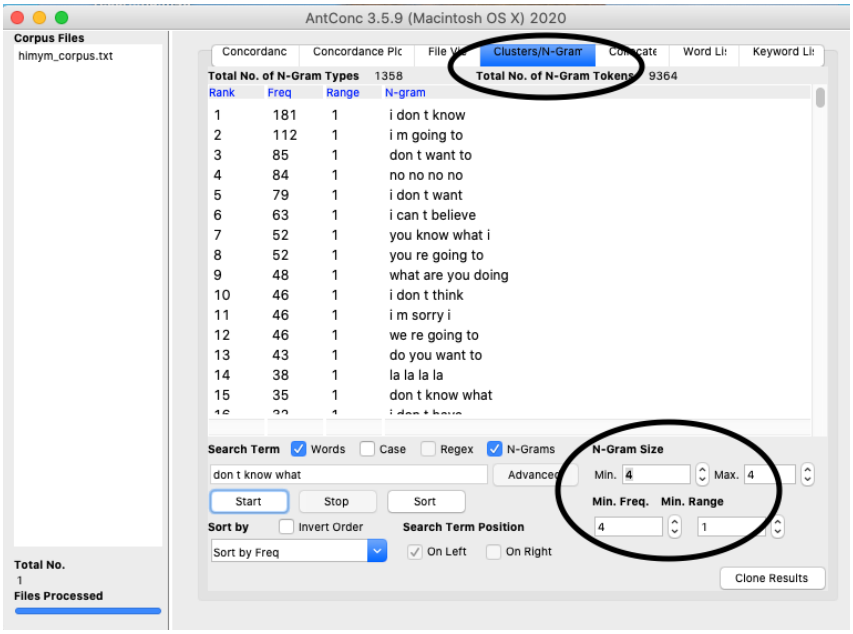

Figure 1 is an example of how I set up the AntConc program to search for the most frequent four-word lexical bundles in the HIMYM corpus as well as in the Santa Barbara Corpus. The main function of AntConc used for this research is Cluster/N-Grams. Starting from this, for example, it is possible to get an indication of the amount and type of lexical bundles in the corpus. Firstly, to perform the analysis, it is necessary to have a text in .txt format in order to enter it into the program and be able to search for terms. After that is important to set the different search criteria needed for the study. In this case, I chose to set the minimum frequency criterion of four times and the minimum and maximum size of lexical bundles of four words. It can be noticed how different

results will be achieved in each corpus. This type of technology makes it possible to speed up the whole investigation process, since the operation is instantaneous.

3.1.3 Procedure

As already highlighted, the first digital platform involved was Sketch engine in order to collect each conversation and consequently create a new corpus. Once I collected the two corpora available and especially understood how to use the technological tools for this corpus analysis, I started analysing and finding lexical bundles for each corpus with the AntConc program. After the corpus (in .txt) was entered into AntConc, the number and frequency of lexical bundles found could be checked and reported using the Clusters/N-Grams tool as shown in Figure 1. By fixing a minimum frequency (four times) and a precise size (three-five words/four words) of the lexical bundles that I decided to search, I then reported the results and immediately chose to focus on four frequently used lexical bundles (four-word strings). After finding these multi-words sequences in both corpora, I compared them; the comparisons were made between the expert's corpus (Santa Barbara Corpus) and my corpus (HIMYM Corpus) to find out frequencies and functions of the most frequent lexical bundles. Given the fact that the two corpora used were of different sizes – one including 249,000 words while the other 190,000 – the problem was how to relate the results for both corpora in order to give proper and accountable evidence of my investigation.

Table 1. Number of words in the corpora analysed

	Santa Barbara Corpus of Spoken American English	**HIMYM Corpus**
Size in words	249,000	190,274

As can be seen in Table 1, a significant difference was found in terms of the length of the corpora. When this difference was examined, the results had to be "normalised" using a specific formula. The frequencies per million words and per texts were calculated in order to compare the standardized findings between the corpora. To do so, the results were *normalised* to a common base using a specific formula which assumes that the raw frequency is divided by the total number of words in the corpus section and that the results is multiplied by one thousand. For example, if the frequency of a given

word is 160 in the 249,000 words corpus and 50 in the 190,000 words corpus, the normalised frequencies are:

- 160 ÷ 249,000 * 1,000 = 0.64 occurrences per thousand words
- 50 ÷ 190,000 * 1,000 = 0.26 occurrences per thousand words

After establishing the two balanced corpora and identifying the "normalised" frequency of the lexical bundles in each corpus, true comparison of the two corpora could be made. Finally, I selected four three-word lexical bundles that were most frequent in my research in order to describe the similarities and differences between the two corpora employed, regarding the use and function of these lexical bundles in discourse. In fact, I provided examples of the use of the lexical bundle found for each corpus so that their discourse function, following the categorization of Biber et al. (1999), was clear and could be easily analysed.

3.2 Results and Discussion

3.2.1 Data collection

First of all, the overall number of three to five-word lexical bundles in the Santa Barbara Corpus of Spoken American English and HIMYM Corpus were identified. Three-to-five-word sequences were considered the most appropriate length to generate sufficient bundles, while with two-word bundles it can be difficult to identify functions. By entering the separate texts into the AntConc software, a first result could be seen regarding the presence of N-grams in each corpus.

Table 2. Total frequencies of three to five-word lexical bundles

	No. of N-Gram Tokens	*Corpus size*	*Percentage*
Santa Barbara Corpus	965,238	249,000	26%
HIMYM Corpus	601,710	190,274	31%

Table 2 demonstrates the overall frequencies of lexical bundles in the separate corpora. As can be seen in this table, the frequency of tokens in the two corpora was calculated also in percentage in to make a comparison. The percentage of N-grams found in the two analysed corpora was derived, and AntConc revealed that there is a higher three-to-five N-grams presence in HIMYM Corpus (31%) rather than in the Santa Barbara Corpus (26%). Despite this, however, it is not a relevant difference and makes it clear

that the presence of linguistic elements of conversation are used in the scripted speech of the TV series (such as HIMYM) as well as in the spoken American English of real-life conversation.

3.2.2 Top four-word lexical bundles

Going into more detail, I propose a ranking (top five) of the most frequent four-word clusters in the separate corpora in order to understand the type of lexical bundles that are used in the language of television compared to those found in American spoken English. This research only focused on four-word bundles as previous literature indicated that three and four-word strings are the most researched size and had been used in many related studies. Lexical bundles that included repetitions (*yeah yeah yeah*) and words that clearly demonstrated hesitations (*er, erm*) were excluded, as although these clearly have a pragmatic function, they are not lexical items that would be considered.

Table 4-5. The most frequent four-word lexical bundles in the corpora

	Bundle type	***Frequency***	***Normalised Frequency PKW***
Santa Barbara Corpus *Table 4.*	I don't know	370	1.5
	I don't think	84	0.34
	don't know if	56	0.22
	that's what I	53	0.21
	don't know what	50	0.20
	you know what I	42	0.17
	you don't have	39	0.16
	don't have to	37	0.15
	I don't have	37	0.15
	don't know how	36	0.14
HIMYM Corpus *Table 5.*	I don't know	181	0.9
	I'm going to	112	0.58
	don't want to	85	0.44
	I don't want	79	0.41
	I can't believe	63	0.33
	you know what I	52	0.27
	you're going to	52	0.27
	what are you doing	48	0.25
	I don't think	46	0.24

According to Table 4 at the top of the list of the Santa Barbara Corpus of spoken American English, *I don't know* occurs at a comparatively high frequency level and, indeed, it is employed 370 times (normalised frequency of 1.5 per thousand words);

similarly, it can be seen that the same lexical bundle also tops the ranking of the HIMYM Corpus. On the other hand, the Table 5 shows that the four-word bundle *I don't think*, although is frequently used in the Santa Barbara Corpus (84 times, 0.34 occurrences per thousand words), in the HIMYM Corpus is not even in the Table above, meaning that is not typical of television language (46 times, 0.24 occurrences per thousand words). Moreover, it is interesting to note that the expression I can't believe" earns a place in the top five most frequent lexical bundles of the HIMYM corpus (a frequency of 0.33 per thousand words), while in the Santa Barbara Corpus appears just 11 times (0.04), probably suggesting that in spoken American English there is a preference for using less articulate and more colloquial expressions (e.g. *wow, really*) over these prefabricated and more informal expressions. Comparing these with the most frequent bundles in Santa Barbara Corpus there are just three sequences that are perfectly equals (i.e. *I don't know, I don't think* and *you know what I*), however all the lexical bundles found in the corpora have similar discourse functions as we will see in the next paragraph.

3.2.3 Functional analysis of four lexical bundles

As we have seen in the first chapter, *lexical bundles* are classified in terms of their discourse functions (Biber *et al*. 1999). The main functional types are *stance bundles*, *discourse organizers* and *referential bundles* and relevant sub-categories. Despite the general categorization, it is important to focus on *stance bundles*, which are more frequent in spoken English.

Table 6. Stance bundles (Biber et al., 2004, pp.384-388)

Functional categories of lexical bundles	Examples
1. Stance Expressions	
• Epistemic stance	
➢ Personal	*I don't know if, I think it was*
➢ Impersonal	*are more likely to, the fact that the*
• Modality/attitudinal stance	
➢ Desire	
▪ Personal	*I don't want to, what do you want*
➢ Obligation/directive	
▪ Personal	*you need to know, I want you to*
▪ Impersonal	*it is necessary to, it is important to*
➢ Intention/prediction	
▪ Personal	*I am going to, are we going to*
▪ Impersonal	it'*s going to be, are going to be*
➢ Ability	
▪ Personal	*to come up with, to be able to*
▪ Impersonal	*it is possible to, can be used to*

Table 6 lists stance expressions which normally reflect the speaker or writer's attitudes with respect to a given speech or his/her degree of certainty. This category, according to Biber et al. has two major sub-categories that relates to the more specific functions or contextual meaning, namely, epistemic and attitudinal/modality. Epistemic stance bundles express the speaker's degree of certainty or uncertainty. Attitudinal/modality stance bundles reflect the speaker's attitudes towards the events that happen in speech: desire (e.g., don't want to), obligation/directive (e.g., I want you to), intention/prediction (e.g., is going to be), and ability (e.g., to be able to). The stance bundles can be further categorized as personal or impersonal. Personal stance bundles are directly referred to the speaker; impersonal stance bundles, on the other hand, convey the same meaning but they do not include a direct reference to the speaker.

Table 7. Discourse functions of lexical bundles found

Stance expressions	
1.Epistemic stance	
-Personal	I don't know, I don't think, don't know if, don't know
-Impersonal	what, you know what I, don't know how
2.Modality/Attitudinal stance	
-Desire	don't want to, I don't want,
-Obligation/directive	you don't have, I don't have, don't have to
-Intention/prediction	you're going to, I'm going to, what are your doing

As can be seen in table 7, most of the lexical bundles reported in the previous tables (4-5), belong to the functional category of stance expressions. In this table, the lexical bundles found are divided into their sub-category (epistemic or modality/attitudinal stance), so that we have a clear view of the functions these lexical bundles serve in discourse and in order to have the possibility to examine them in more detail. In this paragraph I will analyse the first four more frequent lexical bundles found in the corpora which are typical of colloquial English.

Table 8. Four-word lexical bundles

LEXICAL BUNDLES	NORMALISED FREQUENCY IN SBCSAE	NORMALISED FREQUENCY IN HIMYM CORPUS
I DON'T KNOW	1.5	0.95
I'M GOING TO	0.08	0.58
I DON'T THINK	0.34	0.24
DON'T WANT TO	0.07	0.44

Table 8 shows the four lexical bundles I decided to focus on: "I don't know", "I'm going to", "I don't think" and "don't want to", clearly resemble each other structurally and most importantly some of them have similar discourse function. Once calculated the normalised frequency of the lexical bundles chosen, I will go into a detailed analysis of each.

3.2.3.1 Lexical bundle "I don't know"

The expression "I don't know" has a higher frequency in the Santa Barbara Corpus than in the HIMYM Corpus. In the first corpus it occurs 370 times and the normalised number is 1.5 occurrences per thousand words, whereas in the TV series corpus appears 181 times, namely 0.95 occurrences per thousand words. Despite this, the difference is not relevant since *I don't know* is the most frequent lexical bundle used in both Santa Barbara Corpus and HIMYM Corpus. As far as its contextual function is concerned, it appears to be slightly different in the corpora analysed, even though they belong to the same functional category: stance bundles.

(1) MARSHALL: You okay?
TED: Sure. Why?
MARSHALL: Oh, *I don't know.* Girl of your dreams dating a billionaire.

(2) LENORE: What what what is this?
ALINA: *I don't know.*

The example above (1) reports a conversation between two protagonists of the TV series How I Met Your Mother (Marshall and Ted) in a rather sarcastic dialogue. Here the lexical bundle *I don't know* has a discourse-marking function and allows thinking time to hedge the utterance and to add further vagueness to a suggestion. On the other

hand, in the example 2 it can be seen how expression *I don't know* in Santa Barbara Corpus is used in the sense of lacking knowledge or certainty about a thing or fact. Here (2) the lexical bundle is a stance bundle in the sub-category of epistemic (personal) stance that include personal expression and an utterance of uncertainty.

3.2.3.2 Lexical bundle "I don't think"

In Santa Barbara Corpus the lexical bundle "I don't think" occurs 84 times and the normalised number is 0.34 occurrence per thousand words, meaning that this four-word sequence has a relevant frequency in conversations of American English. However, in the HIMYM corpus it occurs fewer times, marking a normalised frequency of 0.24 per thousand words (46 times). Similar to the lexical bundle "I don't know", mentioned in the previous paragraph, *I don't think* also has the function of a stance bundle and the examples below show how this expression has the same discourse function in the two corpora:

(3) TED: I did!
BARNEY: Yeah, *I don't think* you did.

(4) WALT: *I don't think* he walked up and shook his hand, we're gonna see that in a second.

Here (3), (4), speakers used personal epistemic stance bundles to show that they were not completely sure about what they were saying. The lexical bundle *I don't think,* indeed, shows the degree of knowledge or information of the speaker, who expresses a personal assessment of the issue discussed and makes his/her assumptions. In the example (3) Barney is arguing with the main character of the TV series Ted about something he thinks he did not do; similarly in the second example (4) the speaker expresses his uncertainty about something that is thought to have happened. In both contexts *I don't think* allows the speaker to express an opinion about a fact, so it falls into the functional sub-category of epistemic (personal) stance.

3.2.3.3 Lexical bundle "I'm going to"

I'm going to" occurs very few times in the Santa Barbara Corpus, only 20, namely 0.08 occurrences per thousand words. The situation is different for the use of this lexical bundle in the HIMYM Corpus, where it appears 112 times, marking a normalised frequency of 0.58 per thousand words. Nevertheless, it is important to reflect on the form of this lexical bundle chosen to be used in the different corpora. In fact, *I'm going to* is often replaced by "I'm gonna", particularly in American English, but this is never done in writing. *I'm gonna* is just the informal version that one says to friends and family, whereas *I'm going to* is the polite way of saying it. This could be the reason why in the Santa Barbara Corpus, which reports real-life conversations, the frequency of the lexical bundle *I'm going to* is lower than in the corpus of the TV series How I Met Your Mother. Despite this, it is still possible to examine the function of the multi-words sequence mentioned.

(5) TED: *I'm going to* make a phone call.
ROBIN: I'll preheat the oven.

(6) KEVIN: *I'm going to* take a bite.

Both the examples above, show that the lexical bundle analysed belongs to the functional sub-category of attitudinal/modality, which reflect the speaker's attitudes towards the events. In the first case (5) there is a conversation between two actors from the TV series (HIMYM) who are in an awkward situation, and in particular Ted uses the four-word expression to introduce and to give an indication of what he is going to do (intention) to get out of the uncomfortable situation he is in. Similarly, in the second example (6) the speaker uses *I'm going to* as a stance-modality bundle, in order to communicate precisely his future (personal) intention.

3.2.3.4 Lexical bundle "don't want to"

The lexical bundle "don't want to" has a higher frequency in the HIMYM Corpus than in the Santa Barbara Corpus: a normalised frequency of 0.44 per thousand words in the former (85 times) and of 0.07 per thousand words in the latter (19 times). In one hand, in the TV series the expression uses a verb at the beginning, emphasizing the

importance of the action, on the other hand in the Santa Barbara Corpus the lexical bundle starts with a personal pronoun and then the verb phrase fragment (e.g. *I don't want to*). It is a subtle difference that may alter the results of the analysis on AntConc.

(7) TED: You just feel like everything you say is gonna make things worse.
BARNEY: Exactly. And you know why?
TED: *Don't want to* hurt someone I really care about.

(8) JAMIE: *Don't want to* take that class. So maybe I'll wait.

Examples 7 and 8 report how the lexical bundle *don't want to* is used to reflect the speakers' attitudes towards the conversation and, more precisely, towards a certain topic. Both of these utterances indeed express a desire (modality/attitudinal stance). Speakers here (7), (8) reveal something that they don't want to happen.

3.3 Conclusions

Altogether, there were a frequency of 965,238 N-Grams tokens in the Santa Barbara Corpus and 601,710 in the HIMYM Corpus respectively, implying that lexical bundles are a relevant part of spoken English. Returning to the research questions posed at the beginning of this chapter, the first and the second one ask what the most frequent four-word lexical bundles in the TV series HIMYM and in the Santa Barbara Corpus are. Results shows that *I don't know* is the most frequently used four-word bundle in the HIMYM Corpus and appear 181 times, marking a normalised frequency of 0.95 occurrences per thousand words. Next, it is found the lexical bundle *I'm going to* (0.58 occurrences per thousand words) and straight after *don't want to* (0.44 occurrences per thousand words).

In answer to question two, in the Santa Barbara Corpus it is possible to see that the lexical bundle *I don't know* has the highest frequency, such as in the HIMYM Corpus, occurring 1.5 times per thousand words. Following this is the expression *I don't think*, which has a normalised frequency of 0.34 per thousand words. These results lead to the conclusion that these fixed expressions analysed have a common structure. More precisely, this study highlights how Spoken American English tended to use "verb phrase", although lexical bundles are either dependent clause or noun phrase and

prepositional phrase. Similarly, Biber et al. (2004) discovered that conversation uses more verb phrases and does not normally tend to use bundles with noun and prepositional phrases, as opposed to academic prose. This corpus-based analysis indeed revealed that lexical bundles in the HIMYM Corpus as well as in the Santa Barbara Corpus mainly contained verb phrase fragments and that speakers seem to rely more on these structures in order to express their opinion (e.g., *I don't think, I don't know*).

Finally, question three focuses on the linguistic sequences (of each corpus) in relation to their contextual functions. In fact, it is remarkable that all the lexical bundles, I decided to analyse, nearly belong to the same functional category, namely that of *stance expressions*. An exception can be seen in the example (1) where the actor of the TV series HIMYM makes use of the lexical bundle *I don't know* as a discourse marker to promote the smooth running of the interaction and avoid it breaking down. Despite this, the corpus analysis conducted suggests that stance bundles account for the highest proportion in the spoken genre, aiming to express the speakers' degree of certainty or uncertainty and their wishes or request. In this research, some of the lexical bundles mentioned belong to the attitudinal/modality category, such as *I'm going to*, which expresses an intention, and *don't want to* (desire), which typically includes the pronoun "I" in the Santa Barbara Corpus but in the HIMYM Corpus appears more frequently with the verb at the beginning. On the other hand, lexical bundles *I don't know,* and *I don't think* are placed in the (personal) epistemic stance sub-category and are used in order to index degrees of uncertainty towards the clause to which they are attached. In conclusion, the results show overlap in both form and functions of bundles, but the overall impression created by the most frequent bundles is slightly different. While on one hand, in fact, the structures of the lexical bundle analysed (*I don't know, I don't think, I'm going to, don't want to*) are equal and they belong to the same functional category (stance expressions), on the other hand these expressions convey different discursive intentions (desire, certainty, intention etc.), causing them to belong to different functional sub-categories (e.g. epistemic or attitudinal/modality stance).

Conclusion

Overall, this paper revealed that characteristic spoken *lexical bundles* are also representative of the American sitcom "How I Met Your Mother", although the scripts of TV series are well-formulated conversations, and not spontaneous utterances.

The first part of the corpus-study focused on the frequency of extracted three-to-five-word lexical bundles occurring at least 4 times across the two corpora (i.e. Santa Barbara Corpus, HIMYM Corpus). Subsequently, by comparing the lexical bundles found in AntConc in the Santa Barbara Corpus to the HIMYM Corpus ones, I noted that they are very similar in structure and discourse function. More precisely, as inferred from the corpus-based analysis, the lexical bundle "I don't know" is the most frequently used four-word bundle in both HIMYM Corpus (0.95 occurrences per thousand words) and in Santa Barbara Corpus (1.5 times per thousand words).

In addition to this, lexical bundles relevant for the research conducted were "I don't think", "I'm going to" and "don't want to" which have a common "verb phrase" structure, which is preferred by speakers of spoken American English who rely more on these structures in order to express their opinions. Moreover, significant importance must be given to the functional categorization, by Biber et al (2004), used to analyse the four-word strings in the two corpora. Findings and especially examples reveal that almost all the four more closely analysed lexical bundles belong to the category of stance expressions, in which the speaker can express personal opinions, intentions and predictions. Despite the same discourse function (stance bundles), the fixed expressions mentioned at the end of the last chapter belong to two different functional sub-categories, based on whether they want to express an intention/prediction using *I'm going to* (attitudinal/modality stance), uncertainty with *I don't know* and *I don't think* (personal epistemic stance) or a desire by using the lexical bundle *don't want to* (attitudinal/modality stance).

The results of this corpus-based study led to the conclusion that formulaic language of real-life conversations, more precisely the category of lexical bundles, is included in TV discourse, namely scripted language, as it usually aims to reproduce reality and to reflect the spontaneity of authentic spoken language.

Despite this, my research has some limitations. In fact, a first limitation concerning the "HIMYM Corpus" refers to the fact that not only represents just a small portion of American spoken English in scripted language, but it is also very limiting since, by telling the story of some friends in a specific city (New York), it represents a certain way of speaking typical of that place and culture, which may include expressions or specific elements characteristic of that city, excluding other types of American English. In addition to this, collecting data from a single TV series, rather than various sitcoms, on one hand facilitates an identification of lexical bundles, on the other hand is a limitation as the American sitcom *How I Met Your Mother* and corpus created only represent a part of the scripted language of fiction, as it includes only three seasons (64 episodes) of the series, restricting the analysis. Another point that does not make my analysis exhaustive is the number of lexical bundles analysed in more depth. At the end of the third chapter, I investigated only four frequently used bundles which are not sufficient to represent formulaic language and the entire category of lexical bundles and the presence of authentic language in television language.

Since this dissertation is limited, more in-depth studies should be conducted on the authenticity of the language of television, e.g., analysing various TV series or movies to include and compare other varieties of English, further investigating movies to see whether other genre categories produce different results, and consequently using a wider number of corpora to identify and compare linguistic elements of authentic spoken language. In other words, the results of the present dissertation should be verified in other comparable corpora in such a way as to allow other researchers to check whether the results are consistent or not. Furthermore, I think that future research could extend this personal and restricted analysis to other features of spoken language and spoken conversation.

Bibliography

Al Surmi, M. (2018). *Authenticity and TV Shows: A Multidimensional Analysis Perspective.* TESOL Quarterly, vol. 46, pp. 671-694.
From: https://doi.org/10.1002/tesq.33

Altenberg, B. (1998). *On the phraseology of spoken English: the evidence of recurrent word combinations. In Phraseology: Theory Analysis and Applications, ed. by Anthony Paul Cowie,* pp. 101-122. Oxford: Oxford University Press.

Annual Review of Applied Linguistics (2012), vol.32. Cambridge Core *Special issue on Topics in Formulaic Language.* From: https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/volume/6A7C92B302F8CD1CCE5CABE0DAFE3513

Anthony, L. (2019). *Antconc.* Retrieved from Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. From: http://www.laurenceanthony.net/

Baumgarten, Nicole (2005): *The Secret Agent: Film Dubbing and the Influence of the English Language on German Communicative Preferences – Towards a Model for the Analysis of Language Use in Visual Media.* PhD Dissertation. Hamburg: Universität Hamburg. From:https://doi.org/10.13140/RG.2.2.19258.93126

Bednarek, M. (2018). *Language and television series: A linguistic approach to TV dialogue,* pp. 245. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English,* pp. 990. Harlow, UK: Pearson Education.

Biber, D., & Burges, J. (2000). *Historical Change in the Language Use of Women and Men: Gender Differences in Dramatic Dialogue.* Journal of English Linguistics, pp. 21 - 37. From: https://doi.org/10.1177/00754240022004857

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at...: Lexical bundles in university teaching and textbooks. Applied Linguistics*, 25, pp. 371–405.

Biber, D., and Barbieri, F. (2007). *Lexical bundles in university spoken and written registers.* English for Specific Purposes, pp. 263–286.
From: https://doi.org/10.1016/j.esp.2006.08.003

Biber, D. (2009*). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing*. International Journal of Corpus Linguistics, pp. 275-311. From:https://doi.org/10.1075/ijcl.14.3.08bib

Bubel, C. (2006), *The linguistic Construction of Character Relations in TV Drama: Doing Friendship in Sex and the City.* Unpublished PhD dissertation, Universität des Saarlandes, Saarbrücken (Germany). Available at: Diss_Bubel_publ.pdf (uni-saarland.de)

Butler, C. (2003). *Multi-word sequences and their relevance for recent models of functional grammar.* Functions of Language*,* vol. 10, pp.179-208. From: JOURNAL https://doi.org/10.1075/fol.10.2.03but

Cameron, D. (2001) *Working with Spoken Text*, pp. 26. London: Sage.

Cobb, T. (2018). *From corpus to CALL: The use of technology in teaching and learning formulaic language.* In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective*, pp. 192-211. New York: Taylor & Francis.

Cortes, V. (2004). *Lexical bundles in published and student disciplinary writing: Examples from history and biology.* English for Specific Purposes, vol. 23, pp. 397-423. From:https://doi.org/10.1016/j.esp.2003.12.001

Cortes, V. (2006). *Teaching lexical bundles in the disciplines: An example from a writing intensive history class.* Linguistics and Education, pp. 392. From:https://doi.org/10.1016/j.linged.2007.02.001

Cullen, R., & Kuo, I.-C. (2007). *Spoken grammar and ELT course materials: A missing link?* TESOL Quarterly, pp. 361–386. From: https://doi.org/10.1002/j.1545-7249.2007.tb00063.x

Dose, Stefanie (2012): *Scripted speech in the EFL classroom: The Corpus of American Television Series for teaching spoken English*. In Thomas J.and Boulton A.: *Input, Process and Product: Developments in Teaching and Language Corpora,* pp. 62–80. Brno: Masaryk University Press.

Ellis N, Simpson-Vlach R, Römer U, O'Donnell M, Wul S. (2015). *Learner corpora and formulaic language in second language acquisition research.*
From: LCR_offprint_pp357-378.pdf

Erman, B., & Warren, B. (2000). *The idiom principle and the open choice principle.* Text and Talk, vol. 20. From: https://doi.org/10.1515/text.1.2000.20.1.29

Forsyth, R., & Grabowski, Ł. (2015). *Is there a formula for formulaic language?*
Poznan Studies in Contemporary Linguistics, pp. 7.
From: https://doi.org/10.1515/psicl2015-0019.

Grant, L., & Starks, D. (2001). *Screening appropriate teaching materials: Closings from textbooks and television soap operas.* International Review of Applied Linguistics in Language Teaching, pp. 39–50. From: https://doi.org/10.1515/iral.39.1.39

Győrfi, A. 2017. *Formulaic Language: The Building Block of Aphasic Speech*. Advances in Speech-language pathology, pp. 29.
From: 10.5772/intechopen.70038

Hakuta, K. (1974). *Prefabricated patterns and the emergence of structure in second language acquisition.* Language Learning, pp. 287-297. From: https://doi.org/10.1111/j.1467-1770.1974.tb00509.x

Hatami, S. (2015). *Teaching Formulaic Sequences in the ESL Classroom.* TESOL Journal, vol. 6, pp. 112–129. From: https://doi.org/10.1002/tesj.143

Hilliard, A. (2014). *Spoken grammar and its role in the English language classroom.* English Teaching Forum, pp. 2–13.
From:https://files.eric.ed.gov/fulltext/EJ1050242.pdf

Kimball T. (2018). *An overview: American television series as authentic resources for English language development.* University of North Florida. STEM Journal, vol. 19, pp. 1-19. From: https://doi.org/ 10.16875/stem.2018.19.2.1

Lexical Computing CZ s.r.o. (2022). *Create and search a text corpus | Sketch Engine*. Sketch Engine. https://www.sketchengine.eu/

Liu, D. (2003). *The Most Frequently Used Spoken American English Idioms: A Corpus Analysis and Its Implications.* TESOL Quarterly, pp. 671-700. From: https://doi.org/10.2307/3588217

Meskill, C. (1996). Listening skills development through multimedia. *Journal of Educational Multimedia and Hypermedia,* pp. 179-202.
From: https://dcmp.org/learn/166

Mitterer, H., & McQueen, J. (2009). *Foreign subtitles help but native-language subtitles harm foreign speech perception.* PLOS ONE, pp. 77–85. From: https://doi.org/10.1371/journal.pone.0007785

Mittmann, Brigitta (2006): *With a little help from Friends (and others): Lexico-pragmatic characteristics of original and dubbed film dialogue*. In Levshina, N. (2017). *A Multivariate Study of T/V Forms in European Languages Based on a Parallel Corpus of Film Subtitles.* Research in Language, vol. 15, pp. 153–172.
From: https://doi.org/10.1515/rela-2017-0010

Moon, R. 1998. *Fixed expressions and idioms in English,* pp. 145-6. Oxford: Clarendon Press.

Myles, F., & Cordier, C. (2017). *Formulaic sequences (FS) cannot be an umbrella termi in SLA*. Studies in Second Language Acquisition, pp. 3–28.
From: https://doi.org/10.1017/s027226311600036x
Nattinger J. and DeCarrico J., (1992), *Lexical Phrases and Language Teaching,* Oxford University Press, Oxford. Ial Issues in Applied linguistics, vol 5.
From:https://doi.org/10.5070/L451005177

Paquot, M., & Granger, S. (2012). Formulaic Language in Learner Corpora. Annual Review of Applied Linguistics, pp. 130–149.
From: https://doi.org/10.1017/s0267190512000098

Pawley, A., & Syder, F.H. (1983). *Two puzzles for linguistic theory: nativelike selection and nativelike fluency*. In J. Richards, R. Schmidt (Eds.), *Language and Communication*, pp. 191-226. Longman, London.

Peters, A. M. (1977). L*anguage learning strategies: Does the whole equal the sum of the parts?* Language, 53 (3), pp. 560–73.

Peters, A. M. (1983). *The units of language acquisition.* Cambridge: Cambridge University Press, JSTOR, pp. 360-362. From: https://www.jstor.org/stable/44488572

Plunkett, K. (1990). *The segmentation problem in early language acquisition.* Center for Research in Language Newsletter, pp. 1–17.
From: https://doi.org/10.1017/S0305000900009119

Porter, D., & Roberts, J. (1987). *Authentic listening activities*. In M. Long & J. Richards (Eds.), *Methodology in TESOL*, pp. 177-190. Rowley, MA: Newbury House.

Quaglio, P., & Biber, D. (2006). *The grammar of conversation*. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics,* pp. 692–723. Wiley-Blackwell. From:https://doi.org/10.1002/9780470753002.ch29

Quaglio, Paulo (2009): *Television Dialogue: The Sitcom Friends vs. Natural Conversation*. Amsterdam: John Benjamins.

Rey, J. (2001). *Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966 to 1993*. In S. Conrad & D. Biber (Eds.). *Variation in English: Multi-dimensional studie*s, pp. 138-156. London, England: Longman.

Rezaie, O., Farahani, M. V., & Masoomzadeh, M. (2020). *Lexical bundles in PhD dissertations and master theses: A comparative inquiry*. Bulletin of the Transilvania University of Braşov, pp. 127–152.
From: https://doi.org/10.31926/but.pcs.2020.62.13.3.8

Rings, L. (1986). *Authentic language and authentic conversational texts.* Foreign Language Annals, pp. 203-208.
From:https://doi.org/10.1111/j.19449720.1986.tb02835x

Rogers, C., & Medley, F. (1988). *Language with a purpose: Using authentic materials in the foreign language classroom.* Foreign Language Annals, pp. 467-478. From: https://doi.org/10.1111/j.1944-9720.1988.tb01098.x

Rost, M. (2002). *Teaching and researching listening: Harlow, England: Pearson Education.* Taylor & Francis Group. From: https://doi.org/10.4324/9781315833705

Schmitt, N., and Carter, R. (2004). *Formulaic sequences in action: an introduction.* In Schmitt N. *Formulaic Sequences: Acquisition, Processing and Use.* Amsterdam: John Benjamins, pp. 1–22. From: https://doi.org/10.1075/lllt.9.02sch

Schmitt, N. (2010). *Researching Vocabulary – A Vocabulary Research Manual.* Basingstoke: Palgrave Macmillan, pp.117.
From: http://dx.doi.org/10.1057/9780230293977

Scott, M. (1997): WordSmith Tools Manual. Oxford: Oxford University Press.
From: https://lexically.net/downloads/version6/wordsmith6.pdf

Simpson-Vlach, R. & Ellis, N. C. (2010). *An Academic Formulas List: New Methods in Phraseology Research.* Applied Linguistics, vol. 31, pp. 487-512.
From:https://doi.org/10.1093/applin/amp058

Sinclair, J. (1991). *Corpus, Concordance, Collocation,* pp. 110. Oxford: Oxford University Press.

Sinclair, J., Renouf, A. (1991). *Collocational frameworks in Englis*h. In Aijmer, K; Altenberg, B (eds*) English corpus linguistics: Studies in honour of Jan Svartvik,* pp. 128. London: Longman.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. Corpus Linguistics and Linguistic Theory, pp. 11. From: https://doi.org/10.1515/cllt-2014-0016

Sockett, G. (2011). *From the cultural hegemony of English to online informal learning: Cluster frequency as an indicator of relevance in authentic documents.* Groupe d'Étude et de Recherche en Anglais de Spécialité, pp. 5-20.
From: https://doi.org/10.4000/asp.2469

Stubbs, M. and Barth, I. (2003): *Using recurrent phrases as text-type discrimina-tors: a quantitative method and some findings.* Functions of Language, pp. 61-104
From: https://doi.org/10.1075/fol.10.1.04stu

Tannen D., and Lakoff R. (1994). *Conversational Strategy and Metastrategy in Bergman.* In Tannen D. *Gender and Discourse,* pp.137-73. New York: Oxford University Press.

Van Lancker Sidtis, D. (2012). *Formulaic Language and Language Disorders*. Annual Review of Applied Linguistics, pp. 62–80.
From: https://doi.org/10.1017/s0267190512000104.

Wang, Y. (2012). *Learning L2 vocabulary with American TV drama: From the learners' perspective.* English Language Teaching, vol. 5.
From: https://files.eric.ed.gov/fulltext/EJ1079740.pdf

Wang, Y. (2018). *Formulaic sequences signalling discourse organisation in ELF academic lectures: a disciplinary perspective.* Journal of English as a Lingua Franca, pp. 355-376. From: https://doi.org/10.1515/jelf-2018-0017

Werner, V. (2020). *TV Discourse, Grammaticality, and Language Awareness. Teaching English as a Second or Foreign Language: TESL-EJ*, pp. 1–23. From: https://doi.org/10.20378/irb-49030

Wood, D. A. (2006). *Uses and Functions of Formulaic Sequences in Second Language Speech: An Exploration of the Foundations of Fluency.* Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes, pp. 13–33. From: https://doi.org/10.3138/cmlr.63.1.13

Wood, D. (2009). *Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study*. Canadian Journal of Applied Linguistics, vol. 12, pp. 39-57.
From: https://journals.lib.unb.ca/index.php/CJAL/article/view/19898

Wood, D. A. (2015*). Fundamentals of Formulaic Language: An Introduction*, pp. 1-18. Bloomsbury Academic.

Wray, A. (1999). Formulaic language in learners and native speakers. Language Teaching, Cambridge University Press, pp. 213–231.
From:https://doi.org/10.1017/s0261444800014154

Wray, A., & Perkins, M. R. (2000). *The functions of formulaic language: An integrated model.* Language & Communication, vol. 20, pp. 1–28.
From: https://doi.org/10.1016/S0271-5309(99)00015-4

Wray, A. (2002). *Formulaic language and the lexicon*, pp. 1-43. Cambridge, UK: Cambridge University Press.

Zhang, L., & Lu, P. (2017). Lexical Chunks Formulaic Sequences and Yukuai: *Study of Terms and Definitions of English Multiword Units.* English Language and Literature Studies, pp. 74. From:https://doi.org/10.5539/ells.v7n1p74

Summary in Italian

Questa tesi analizza la frequenza di elementi linguistici chiamati “lexical bundles”, ossia fasci lessicali, nel linguaggio inglese parlato e soprattutto, in quello della serie TV “How I Met Your Mother”, linguaggio pianificato e scritto, o “scripted language”. Tutto ciò al fine di confermare che la frequenza, la tipologia e la funzione discorsiva dei fasci lessicali trovati nella lingua parlata (lingua “autentica”), sono rappresentativi anche della sitcom americana, sebbene i copioni delle serie televisive siano conversazioni ben formulate e non enunciati spontanei. Attraverso un’analisi di due corpora, una collezione di testi in formato elettronico selezionati ed organizzati per facilitare le analisi linguistiche, verranno, nel terzo capitolo, riportati i dati raccolti riguardo a questi fasci lessicali.

Il primo termine preso in considerazione è “formulaic language”, ossia linguaggio stereotipato o formulare, espressione che si riferisce ad “una sequenza, continua o discontinua, di parole o altri elementi che è, o sembra essere, prefabbricata, cioè immagazzinata e recuperata interamente dalla memoria al momento dell'uso” (Wray, 2002: 9). Il linguaggio è infatti in gran parte formulare, ed uno dei punti cardine delle interazioni orali, quindi della lingua parlata. In particolare, questo tipo di linguaggio ha varie funzioni: dalla sua onnipresenza nell’uso della lingua delle persone madrelingua o nelle varie lingue del mondo, alla sua utilità nel produrre un discorso fluente e nel ridurre lo sforzo di elaborazione, in quanto le parole in questo caso sono “immagazzinate come interi nella memoria a lungo termine" (Zhang, Lu, 2017: 74). Questo processo, infatti, aiuta i parlanti a risparmiare lo sforzo mentale per formulare il testo, mantenendo allo stesso tempo un certo ritmo e facilità del discorso. Inoltre, il linguaggio stereotipato gioca un ruolo fondamentale nella competenza comunicativa ed è tipico delle espressioni di scuse (es. *I’m sorry,* ecc.), richieste (es. *could you [please],* ecc.), complimenti, saluti, ringraziamenti e condoglianze, oppure è usato per creare la struttura del discorso (es. *one the other hand, as a result,* ecc.) e per la gestione della conversazione (es. *you know, I mean,* ecc.)

Queste espressioni “prefabbricate”, infine, possono essere suddivise in diverse categorie, tra cui ad esempio: idiomi, verbi frasali (phrasal verbs), collocazioni e fasci lessicali (lexical bundles). Tra queste categorie appena elencate, la mia tesi si sofferma

sui fasci lessicali, il cui termine, “lexical bundles”, viene introdotto per la prima volta da Biber et al. (1999) nel "Longman Grammar of Spoken and Written English" riferendosi a "sequenze di forme di parole che vanno comunemente insieme nel discorso naturale" (Biber et al., 1999: 990). Nel 1999, sempre Biber, propone la prima classificazione strutturale, che dipende sia dalle strutture che dalle funzioni dei fasci lessicali, in cui si individuano tre gruppi principali con le relative sub-categorie. La prima è quella che ritroveremo poi nell’analisi del terzo capitolo ed è quella delle “stance expressions”, ovvero le espressioni di posizione che vengono utilizzate per esprimere atteggiamenti, valutazioni ed opinioni di certezza o incertezza; a seguire ci sono i “discourse organizing bundles”, i fasci organizzativi del discorso; ed infine, i “referential bundles”, o fasci referenziali.

Molto importante per questa tesi è soffermarsi sui fasci lessicali nei corpora, su cui si baserà l’analisi. Queste combinazioni linguistiche di due o più parole vengono identificate in un corpus di linguaggio naturale per mezzo di programmi software di analisi linguistica. In realtà, come vedremo nell’ultimo capitolo, i fasci lessicali vengono ricercati con metodi particolari che si basano su lunghezze e tassi di frequenza preselezionati e che si concentrano sulla funzione e sulla frequenza. Si nota come, per un'analisi delle sequenze formulari Biber e colleghi si sono basati sulla "frequenza" (i fasci lessicali dovrebbero presentarsi più volte in un milione di parole per essere considerati) come "criterio di identificazione", ma anche sulla lunghezza di essi (Biber & Conrad, 1999: 181-189). Questo metodo di ricerca mostra che le espressioni frequenti e i fasci lessicali più lunghi di due parole tendono a essere più impiegati (frequenza) per le funzioni discorsive, e quindi più essenziali per il discorso; inoltre, Biber identifica i fasci lessicali di 3 e 4 parole come i più utilizzati nella natura del linguaggio proprio per questo saranno protagonisti della mia analisi (Biber, 1999).

Infine, tornando alle funzioni che hanno queste combinazioni di parole, è necessario sottolineare che, nel linguaggio parlato, la padronanza e l'uso di quest’ultime possono ridurre lo sforzo dell'ascoltatore nella conversazione, facilitando così l'efficienza e l’efficacia della comunicazione, permettendo all'oratore di essere sicuro che ciò che sta dicendo sarà compreso. In sostanza, il linguaggio stereotipato e, più precisamente, i fasci lessicali sono fondamentali per l'uso accurato, appropriato e complessivamente fluente del linguaggio, soprattutto quello parlato.

È proprio sul linguaggio parlato che si concentrerà il secondo capitolo, dando inizialmente una definizione di “autenticità”, e successivamente riflettendo su quanto sia autentico il linguaggio televisivo, più precisamente quello delle serie TV.

L’aggettivo “autentico” si riferisce all’idea di realismo linguistico, in quanto questa tipologia di linguaggio riflette la realtà. Un chiaro esempio di autenticità del linguaggio è la conversazione, che si svolge in un contesto sociale ed è un discorso spontaneo. L’autentico infatti esprime il linguaggio nella vita reale, in contrapposizione al “non autentico”, che descrive l’uso del linguaggio nel dialogo televisivo, il quale è inizialmente preparato e scritto. Nel contesto della televisione, infatti, si utilizza il linguaggio scritto e programmato, “scripted language”; di conseguenza, un film o un programma televisivo si riferiscono a un discorso che ha origine nella scrittura, nel senso che si basano su un documento scritto e creato da sceneggiatori (trascrizione), il quale poi però nella scena diventa linguaggio parlato. Questo dialogo sceneggiato, o scambio parlato tra due o più persone, diventa quindi una rappresentazione della conversazione faccia a faccia. Nel corso degli anni, infatti, le serie TV si sono evolute per avvicinarsi alla realtà e sono diventate una fonte di materiali autentici, in particolare per l'inglese parlato. Sebbene, quindi, l'uso della lingua nel sistema di intrattenimento (televisivo), sia considerato linguaggio “non autentico", in quanto non spontaneo, questo cattura al meglio le caratteristiche linguistiche della conversazione naturale grazie al "codice del realismo" che si riferisce all'imitazione della realtà, facendo si che il dialogo realistico sia una parte rilevante del linguaggio televisivo moderno (Bubel, 2006). Elementi linguistici come, ad esempio, il linguaggio stereotipato o, meglio, i fasci lessicali compaiono nei dialoghi sceneggiati, rendendo il copione meno simulato.

È proprio per questo motivo che i materiali televisivi parlati (es. film, sitcom) possono offrire agli studenti la possibilità di incontrare un linguaggio realistico (parlato) al di fuori dei confini della classe di lingua, e quindi di un contesto istituzionalizzato. Diversi studi, appunto, hanno concordato sul fatto che l'uso di serie TV (in questo caso americane) ha facilitato l’apprendimento di vocaboli (es. parole e modi di dire) che gli studenti hanno dichiarato di non aver mai incontrato nei loro libri di testo, dove l'errata rappresentazione della lingua parlata e la grammatica colloquiale fanno sì che il loro modo di parlare sia molte volte innaturale e troppo formale.

È da questo punto che parte la mia analisi di “lexical bundles” nella serie TV “How I Met Your Mother” e nelle conversazioni in lingua inglese americana. L’analisi di questa tesi riguarda due corpus specifici: il primo è il “Santa Barbara Corpus of Spoken American English”, il quale rappresenta il linguaggio autentico, composto da circa 250,000 parole; il secondo è stato creato appositamente con il programma Sketch Engine, ed è il “How I Met Your Mother Corpus”, il quale rappresenta invece il linguaggio non autentico, composto da circa 190,000 parole. I risultati dell’analisi rispondono ad una serie di domande che vengono poste precedentemente, nel capitolo, e che hanno la funzione di restringere il campo di ricerca, affinché l’analisi risulti più effettiva. Ci sono tre domande alla base di questa analisi del corpus: “quali sono i fasci lessicali di quattro parole utilizzati più frequentemente nella serie TV HIMYM?”, “quali sono i fasci lessicali di quattro parole utilizzati più frequentemente nelle conversazioni della vita reale (Santa Barbara Corpus of Spoken English)?” e ancora “quali sono le somiglianze e le differenze tra i fasci lessicali di quattro parole presenti nei due corpora, seguendo la distinzione delle funzioni discorsive di Biber (Biber et al. 1999)?”.

Attraverso l’utilizzo di un programma di analisi chiamato AntConc, si può procedere con l’analisi iniziale della frequenza dei fasci lessicali da tre a cinque parole, che sono comparsi almeno 4 volte in ciascun corpus. Complessivamente, è emersa una frequenza di 965.238 elementi nel Santa Barbara Corpus e 601.710 nel HIMYM Corpus, il che implica che i fasci lessicali sono una parte rilevante dell'inglese parlato. Successivamente, ho deciso di soffermarmi ed analizzare alcune sequenze di quattro parole presenti, e abbastanza frequenti, in entrambi i corpus scelti (“I don’t know, “I don’t” think”, “I’m going to” e “don’t want to”). Il risultato più rilevante identifica “I don’t know” come la sequenza di quattro parole più frequentemente usata sia nel corpus di HIMYM (frequenza normalizzata di 0,95 occorrenze per mille parole) che nel corpus di Santa Barbara (1,5 occorrenze per mille parole).

I risultati ottenuti inoltre, portano alla conclusione che queste espressioni fisse analizzate hanno una struttura comune. Più precisamente, questo studio mette in evidenza come l'inglese (americano) parlato e la conversazione tendano ad utilizzare la "verb phrase" (frase verbale), rivelando che i fasci lessicali nel HIMYM e nel Santa Barbara Corpus contengono principalmente frammenti di frasi verbali e che i parlanti sembrano affidarsi maggiormente a queste strutture per esprimere la propria opinione

(es. *I don't think, I don't know*). Infine, per quanto riguarda le sequenze linguistiche (di ciascun corpus) in relazione alle loro funzioni contestuali, è interessante notare come i quattro fasci lessicali analizzati, appartengano alla stessa categoria funzionale, quella delle "stance expressions", o espressioni di posizione, suggerendo che questa tipologia di fasci lessicali rappresenti la più frequente nel genere parlato, in cui i parlanti mirano ad esprimere il grado di certezza o incertezza, i loro desideri o richieste. Se da un lato, però, la struttura delle sequenze lessicali analizzate sono uguali e appartengono alla stessa categoria funzionale (espressioni di posizione), dall'altro queste espressioni veicolano intenzioni discorsive diverse (desiderio, certezza, intenzione, ecc.), facendole appartenere a diverse sottocategorie funzionali (epistemica o attitudinale/modale).

Ciò che è emerso è che i fasci lessicali ritrovati nelle conversazioni, linguaggio autentico, sono simili per struttura e funzione discorsiva a quelli trovati nella serie TV "How I Met Your Mother", linguaggio non autentico. Questi risultati sono però limitati, in quanto l'analisi si concentra solo su pochi fasci lessicali, e il corpus preso in considerazione per il linguaggio non autentico comprende solo una sessantina di episodi (HIMYM), che non può essere presa come unica rappresentante di tutte le serie TV per un'analisi completa. Per questo motivo, ulteriori studi potrebbero partire dai risultati di questa tesi e sviluppare ulteriori argomentazioni derivanti dall'analisi di fasci lessicali in altri corpora comparabili, più completi sia dal punto di vista del numero dei testi che del genere di serie TV; in modo tale da permettere ad altri ricercatori di verificare la veridicità e la coerenza delle conclusioni di questa tesi.