

UNIVERSITÀ CATTOLICA DEL SACRO CUORE DI MILANO

Scuola di Dottorato in Scienze Linguistiche e Letterarie

Ciclo XXI

S.S.D.: L-LIN/12

**SPONTANEITY IN AMERICAN ENGLISH:
FACE-TO-FACE AND MOVIE CONVERSATION COMPARED**

Tesi di Dottorato di:
Pierfranca FORCHINI
Matr. N° 3480098

Anno Accademico 2007/2008



UNIVERSITÀ CATTOLICA DEL SACRO CUORE DI MILANO
Scuola di Dottorato in Scienze Linguistiche e Letterarie
Ciclo XXI
S.S.D.: L-LIN/12

**SPONTANEITY IN AMERICAN ENGLISH:
FACE-TO-FACE AND MOVIE CONVERSATION COMPARED**

Tesi di Dottorato di:
Pierfranca FORCHINI
Matr. N° 3480098

Coordinatore: Chiar.ma Prof.ssa Serena VITALE

Anno Accademico 2007/2008

To Lucia, Stefy and Zoe,
To my Karate Teacher, Nicola Ragno,
To my *Nonna Antonia*,
"Amazing Graces... bright shining as the sun..."

To my Mom

TABLE OF CONTENTS

ACKNOWLEDGMENTS	6
INTRODUCTION.....	8
CHAPTER 1. UNDERLYING PRINCIPLES AND METHODOLOGY.....	12
1.1 INTERDEPENDENCE OF LANGUAGE STRATA AND MEANING AS FUNCTION IN CONTEXT.....	12
1.1.1 <i>The Idiom Principle</i>	15
1.1.2 <i>Priming</i>	20
1.1.3 <i>Implications of a Theory of Context</i>	22
1.2 AUTHENTIC DATA AND CORPORA.....	25
1.2.1 <i>Advantages of Authentic-Data Oriented Analyses</i>	26
1.2.1.1 Authentic Data vs. Intuition.....	26
1.2.1.2 Authentic Data, Quantitative and Qualitative Analyses.....	28
1.2.1.3 Authentic Data as Empirical Data	29
1.2.1.4 Authentic Data and Electronic Processing.....	30
1.2.2 <i>Corpora Drawbacks and Possible Solutions</i>	31
1.3 UNITS OF ANALYSIS AND TOOLS.....	33
1.3.1 <i>Multi-Dimensional and Micro Analyses</i>	34
1.3.2 <i>The Longman Spoken American Corpus</i>	41
1.3.3 <i>The American Movie Corpus</i>	42
1.3.3.1 Corpus Building Criteria	45
1.3.3.2 Standardization and Transcription Criteria	49
CHAPTER 2. FACE-TO-FACE CONVERSATION	52
2.1 SPOKEN AND WRITTEN LANGUAGE.....	52
2.2 KEY FEATURES OF SPONTANEOUS CONVERSATION	54
2.3 DISCOURSE MARKERS	60
2.3.1 <i>Terminology, Classification, and Approaches</i>	62
2.3.2 <i>General Traits of Discourse Markers</i>	68
2.3.3 <i>Focus on You Know: a Functional Categorization</i>	74
2.4 SUMMARY OF THE PRESENT FRAMEWORK.....	77
CHAPTER 3. MOVIE CONVERSATION	80
3.1 MULTIPLICITY OF CHANNELS, CODES, AND MESSAGES	82
3.2 FICTITIOUSNESS AND SPONTANEITY	84
3.2.1 <i>Non-Spontaneous Spoken Conversation</i>	84
3.2.1.1 Movie Constraints.....	86
3.2.2 <i>Sounding Spontaneous: the Role of Language and the Movie Scriptwriter</i>	88
3.2.2.1 Linguistic Features	93
3.2.2.2 Discourse Markers: the Case of <i>You Know</i>	96
CHAPTER 4. MULTI-DIMENSIONAL ANALYSIS	102
4.1 FACE-TO-FACE AND MOVIE CONVERSATION COMPARED.....	104
4.1.1 <i>Dimension 1: Informational vs. Involved Production</i>	106
4.1.2 <i>Dimension 2: Narrative vs. Non-Narrative Concerns</i>	111
4.1.3 <i>Dimension 3: Explicit vs. Situation-Dependent Reference</i>	114
4.1.4 <i>Dimension 4: Overt Expression of Persuasion</i>	117
4.1.5 <i>Dimension 5: Abstract vs. Non-Abstract Information</i>	119
4.2 FACE-TO-FACE CONVERSATION AND MOVIE GENRE.....	121
4.3 DISCUSSION OF THE MULTI-DIMENSIONAL RESULTS.....	126

CHAPTER 5. MICRO-ANALYSIS.....	130
5.1 FREQUENCY AND PLOT ANALYSIS OF YOU KNOW	130
5.2 THE DISCOURSE MARKER YOU KNOW.....	134
5.2.1 <i>Turn Positon</i>	137
5.2.2 <i>Functions</i>	137
5.2.2.1 Telling Function.....	138
5.2.2.2 Other Pragmatic Functions	146
5.2.2.3 Functions of <i>You Know</i> within its Turn Position.....	152
5.2.3 <i>Part of a Bigger Cluster?</i>	154
5.2.4 <i>Comedies vs. Non-Comedies: a Matter of Genre?</i>	156
5.3 DISCUSSION OF THE MICRO-ANALYSIS RESULTS.....	159
CONCLUSIONS.....	162
APPENDICES	170
REFERENCES	178

Acknowledgments

This dissertation would not have come into being without precious help and support from many people. First and foremost, I would like to thank my supervisor, Prof. Margherita Ulrych, to whom I am extremely grateful not only for her guidance and encouragement, but also for her trust in me and for involving me in research projects and teaching activities which allowed me to broaden my perspective and gain further experience on many fronts.

I would like to express my profound gratitude to the members of the Department of *Scienze Linguistiche e Letterature Straniere* at *Università Cattolica del Sacro Cuore* (Milan, Italy) for providing me with an intellectually stimulating environment. Special thanks are due to Prof. Luisa Camaiora, Dean of the Faculty of *Scienze Linguistiche e Letterature Straniere*, Prof. Bona Cambiaghi, Head of my Department, and to Prof. Serena Vitale, Coordinator of my Doctoral School. My heartfelt appreciation goes out to Prof. Giovanni Gobber for his PhD (and impromptu corridor) classes and to those working in English linguistics under the supervision of Prof. Margherita Ulrych, in particular to Prof. Maria Luisa Maggioni, Prof. Frances Lonergan, Amanda Murphy, Costanza Cucchi and Sarah Bigi for being extraordinarily encouraging and supportive. Thanks also to Simona Galbusera, the people working in *Necchi 9* and *Morozzo della Rocca*, to my students, and to my PhD mates, in particular to Chiara, Michela, Monica, Sara, and Erica, who made it really unique.

I would like to thank *Northern Arizona University* (USA). I am immensely grateful to Prof. Douglas Biber and to Prof. Randi Reppen for inviting me there, for giving me free access to NAU, software and corpora, and for offering inestimable time, collaboration, and friendship. Prof. Douglas Biber's ideas, methodology and meticulous comments had a major influence on this dissertation, and Prof. Randi Reppen gave me incredible support and consideration; this is probably because "people cluster as words do" ("you know I mean it"). I am also grateful to Nicole Tracy Ventura, whose company and feedback about my American English made my being there even more special.

My sincere thanks goes to Prof. Noam Chomsky, who, despite the non-chomskian slant of the present research, was ready to answer all my questions and doubts.

I am also very thankful to Prof. Roberta Facchinetti, who might not remember it, but who (15 long years ago) first introduced me to English linguistics and just made me love it. Similarly, I would like to thank Prof. Maurizio Gotti, Prof. Maria Pavesi and Prof. Silvia Brutti

for keeping my interest alive. My sincere thanks also to Prof. John Sinclair, who is no longer with us, but whose ideas and personality are still greatly illuminating.

After so many years, I express my everlasting gratitude to Susan Ann Foster, my first American teacher and one of my greatest friends.

I would like to thank my friends wholeheartedly for dearly supporting me when in need, and for putting up with all my ups and downs, with all my “sorry really can’t today”, and with my never coming “next weekend”. My warmest thanks to Alessandro, Paola, Anna, Helen, Rosemarie, Roby, Irene, Luisa, Silvia, Serena, Damiano, Adriana, Vale, Elly, Lea and Myrna. A very special thanks goes to my high school best mate, Monica, who strongly encouraged me to pursue my *PhDream* and to Sr. Angelica, who never forgets me in her prayers.

My heartfelt thanks to *Mamma Ragno*, Ricky (my Karate-Brother), *Anshin-Kai* (especially to Enzo, Diego, Emanuele and Leo), my Karate students (in particular to Manuel and Miki), Mandy, Marino, Alessandro, Jodie and Kevin, who are *just family*.

My deepest ever felt gratitude and thanks to *My Sensei*, Nicola Ragno, who is an everlasting presence along *The Way* and will always be, no matter *where the wind will blow*, and to my extraordinarily exceptional mom and grannie, to whom I owe my degrees and my medals, but above all, to whom I owe my life, my values and the most unconditional love and support.

Introduction

The present research fits into studies of spoken language, one of the foci of current empirical linguistics (cf. Biber *et al.* 1999, McCarthy 1999, Biber and Finegan 2001a, Reppen 2001a, McCarthy 2003, Carter and McCarthy 2006). In detail, it investigates face-to-face and movie conversation, two conversational domains which are usually considered to differ in terms of spontaneity. Face-to-face conversation, indeed, is usually defined as spontaneous because it takes place in real time, is not edited (Chafe 1982, McCarthy 2003, Miller 2006), draws heavily on implicit meaning (since the context is often shared by the participants), and consequently lacks semantic (Bercelli 1999) and grammatical elaboration (Halliday 1985, Biber *et al.* 1999). *Normal dysfluency* (Biber *et al.* 1999:1048) and *fragmented language* (Chafe 1982:39) phenomena, such as repetitions, pauses, and hesitation (Tannen 1982, Bazzanella 1999, Halliday 2005), well illustrate its unplanned, spontaneous nature. On the other hand, movie conversation is usually defined as non-spontaneous in that, by being artificially designed to sound like authentic language, it lacks the spontaneous traits which are typical of face-to-face conversation. Consequently, because of this careful planning, movie conversation is usually described as not being representative of the general usage of conversation (Sinclair 2004b:80).

One of the problems of studying speech is the difficulty to collect data. Thus, given the increasing insistence on authenticity and the complications involved in gathering spoken material, if it could be shown that the conversational domains being examined here display similar linguistic features, it would then be justifiable to use movie data as a potential source for the study of the spoken language, and consequently for spoken language teaching and learning.

The main aim of the research is to collect empirical evidence of the linguistic similarities or differences between face-to-face and movie conversation. The significance of this partly derives from the fact that there are few empirical studies that specifically compare these two domains. Not many scholars have written on actual movie dialogs, and many studies have focused on issues of dubbing or sub-titling movies into other languages, comparing the original and dubbed or subtitled versions (Baccolini and Bollettieri Bosinelli 1994; Pavesi 1994, 2005; Bollettieri Bosinelli 1998; Pavesi and Malinverno 2000; Taylor 2000a; Gottlieb and Gambier 2001; Bruti and Perego 2005; Bruti 2006), rather than comparing movie

language to face-to-face conversation. Secondly, a considerable amount of work has been carried out on movie scripts found on the web (Taylor 1999, Taylor and Baldry 2004), rather than on transcribed movie dialogs. Thirdly, some strongly-worded claims about the non-spontaneity of movie language have been based on intuition, rather than on empirical evidence: Sinclair, for instance, without providing data, maintains that movie language is “not likely to be representative of the general usage of conversation” in that its distinctive features do not “truly reflect natural conversation” (Sinclair 2004b:80). Lastly, apart from some studies on TV series such as *Star Trek* (Rey 2001) and *Friends* (Quaglio 2004), there are no studies of movie language that apply Biber’s (1988) Multi-Dimensional analysis approach, which has proved to be reliable as an empirical method of describing the linguistic characteristics of texts.

First of all, then, the present work addresses the following research question: at a macro-level, to what extent do face-to-face and movie conversation differ or resemble each other? At a micro-level, instead, focus is given to one element, the lexical bundle *you know*, which has a special status in speech (Crystal 1988); indeed, since it is very frequent in conversation (cf. Kennedy 1998, Biber *et al.* 1999), *you know* is claimed to be part of the core spoken language (McCarthy 1999, Erman 2001). The other research questions regard the presence of *you know*: is it equally frequent in movie language? What are its pragmatic functions in both face-to-face conversation and movie language, and do these functions vary according to its position in the turn? The influence of movie genre on this difference or resemblance is also discussed.

To answer these questions, empirical data from an existing spoken American English corpus (i.e. the *Longman Spoken American Corpus*) were analyzed, and a new corpus of American movie conversation was purposely built, transcribed, explored and compared with the spoken corpus. First, Biber’s (1988) Multi-Dimensional analysis approach, which applies multivariate statistical techniques by observing and analyzing more than one statistical variable at a time, was applied to these data. Then, the occurrences of *you know* were functionally investigated in context. The analyses shed light on both the general features of the two conversational domains, and on the specific behavior of *you know* in them.

Conceptually, the work is divided into two main parts: the first provides the theoretical background (Chapters 1, 2, and 3), and the second presents the practical analyses (Chapters 4 and 5). In detail, Chapter 1 outlines the functional, descriptive, corpus linguistic

approach adopted and describes the precise methodology applied to the data, in terms of units of analysis and tools. The purpose of this chapter is to illustrate the relevance to linguistic research of the interdependence of the different levels of language, the notion of meaning as function in context, and corpora. Chapters 2 and 3, instead, address the two conversational domains in question: face-to-face and movie conversation. In particular, Chapter 2 illustrates the reasons why spoken language is a relatively new field of research, to present a taxonomy of the spoken domain, and to provide an overview of the key features of spontaneous conversation. Given the central relevance of discourse markers to speech, it then describes typical ways of classifying them, and discusses their common traits. Especially, it focuses on the categorization of *you know* from a functional point of view, and on the functions it performs according to its position in the turn. Chapter 3, on the other hand, explores the features of movie conversation, a type of speech which is typically described as non-spontaneous, prefabricated and written to imitate authentic language (cf. Sinclair 2004b, Taylor 1999, Rossi 2003, Pavesi 2005). The aim of the chapter is to illustrate the multi-modal nature of movies and the need for the co-presence of fictitious (non-spontaneous) and spontaneous traits in them.

The data analysis, which was made possible especially by the collaboration and support of Prof. Douglas Biber at *Northern Arizona University*, is divided into two chapters. Chapter 4 investigates the two domains (face-to-face and movie conversation) in quantitative and qualitative terms, applying Biber's Multi-Dimensional Analysis approach. According to this methodology, the domains are compared through computerized analyses which reduce a large number of linguistic variables (such as the occurrences of nouns, first person pronouns/possessives, second person pronouns/possessives, wh pronouns, prepositions, verbs, suasive verbs, passive verbs + by, and passive post-nominal modifiers, *inter alia*) to a few basic parameters of linguistic variation (Biber 2004), which in turn characterize specific *dimensions*. Thus, groups of features which usually co-occur in texts are, first, counted to have the exact, quantitative characterization of texts (so that texts can be compared very precisely); then, these groups are interpreted functionally (Biber 1988). This methodology is based on the assumption that, since frequently co-occurring linguistic features in texts share at least one communicative function, it is possible to identify single Dimensions which underline each set of co-occurring linguistic features.

Chapter 5 contains the micro-level analysis of the data, *micro* (and *mono*) in that it

concentrates on only one item (i.e. the discourse marker *you know*). This chapter investigates the frequency and functions of *you know* and the extent to which these functions may or may not vary according to its position in the turn.

The dissertation concludes by arguing that, on the basis of the quantitative and qualitative analyses, movie conversation does not, in fact, differ significantly from face-to-face conversation, and can therefore be legitimately used to study spoken language.

The appendices contain the Multi-Dimensional analyses of the linguistic features in the two domains investigated here.

CHAPTER 1. UNDERLYING PRINCIPLES AND METHODOLOGY

Chapter 1 focuses on the principles underlying the research and on the methodology adopted, and is divided into two main parts. The first, Sections 1.1 and 1.2, is theoretical and describes a functional, descriptive, corpus linguistic approach operating within the framework of a contextual and functional theory of meaning, and availing itself of new technologies such as corpora and computers to describe data. The second, Section 1.3, is more practical and provides methodological information about the units of analysis, the tools, and the corpora used for this research.

The aim of Sections 1.1 and 1.2 is to demonstrate the relevance to linguistic analysis of the interdependence of the different levels of language, of the notion of meaning as function in context, and of authentic data retrieved from corpora. More specifically, Section 1.1 describes the interdependence of language strata together with the consequent need to observe language in its entirety and the notion of meaning as function in context, exemplified through concepts such as the Sinclairian *idiom principle* (Section 1.1.1) (Sinclair 1991) and the Hoeyian idea of *priming* (Hoey 2005) (Section 1.1.2); implications connected to this functional framework are also outlined (Section 1.1.3). Section 1.2, instead, explains the reasons for choosing authentic-data oriented analyses (Section 1.2.1), the drawbacks of corpus studies, and possible solutions to them (Section 1.2.2).

The aim of Section 1.3 is to illustrate the observations of the present research and the way they are measured both at a macro- and a micro-level via quantitative and qualitative analyses. In particular, Section 1.3 explains the units of analyses and tools used and how multivariate statistical techniques work (Section 1.3.1); Sections 1.3.2 and 1.3.3 introduce the corpora the present data are retrieved from: the *Longman Spoken American Corpus* (henceforth LSAC) for American face-to-face conversation and the *American Movie Corpus* (AMC) for American movie conversation. Sections 1.3.3.1 and 1.3.3.2 especially focus on the building and transcription criteria of the latter.

1.1 Interdependence of Language Strata and Meaning as Function in Context

The centrality to linguistic analysis of the mutual relations between the different levels of language and of meaning as function in context arise from a tradition based on the pioneering

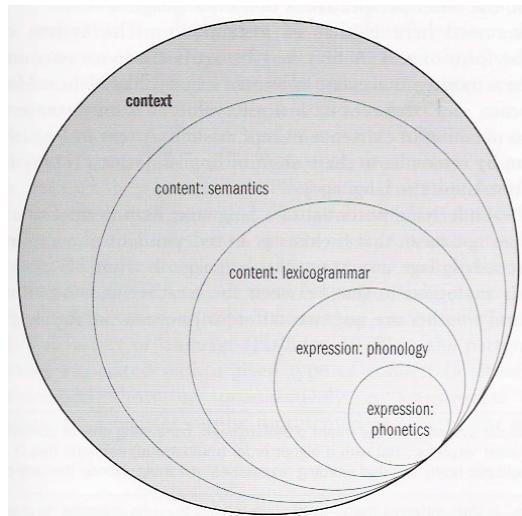
work of John Rupert Firth (Firth 1935a, 1935b, 1951a, 1951b, 1957a, 1957b), which was developed by the so-called new-Firthians – i.e. Michael Halliday and John Sinclair - and, later, by contemporary scholars such as Biber, Francis and Hunston, Stubbs, Hoey, Tognini-Bonelli, *inter alia*.

It is Firth (1951a, 1957b), indeed, who first points out that the language system is based on the mutual relations between the different levels of language which can be identified in context (my emphasis):

It will be noticed that ‘linguistic forms’ are considered to have ‘meanings’ **at the grammatical and lexical levels**, such ‘meanings’ being determined by **interrelations of the forms** in the grammatical system set up for the language. [...] The real units of language are not sounds, or written characters, or meanings: **the real units of language are the relata which these sounds, characters, and meanings represent**. The main thing is not the sounds, characters, and meanings, but **their mutual relations within the chain of speech and within the paradigms of grammar. These relations make up the system of a language**. [...] As a general principle, and as far as possible, the meaning of linguistic forms at the grammatical and lexical levels should be determined with reference to the system of the language and **identified by linguistic context** (Firth 1957b:219-227).

Then, in a similar vein, Halliday (1985a, 1994) and Halliday and Matthiessen (2004:24) further describe language as a complex semiotic system with various levels, which they call *strata*. More specifically, as illustrated in Figure 1, they identify a *stratum of content* and a *stratum of expression*: the *stratum of content* expands into lexicogrammar and semantics. The former is the interface of grammar (i.e. syntax plus morphology) and vocabulary, and represents the *stratum of wording*, while the latter represents the *stratum of meaning*. The *stratum of expression*, on the other hand, expands into phonetics and phonology, the interfacing with the body resources for speech and hearing, and the organization of speech sound into formal structures and system (Halliday and Matthiessen 2004:24-25,587).

Figure 1. Halliday and Matthiessen's (2004:25) stratification



The originality of Halliday's theory, however, does not lie in identifying stratification itself, but in the description of its comprehensive nature. Indeed, this systemic functional perspective considers "language in its entirety, so that whatever is said about one aspect is to be understood always with reference to the total picture" and "what is being said about any one aspect also *contributes* to the total picture" (Halliday and Matthiessen 2004:19-20).

Another key factor of Halliday's systemic functional approach (1985, 1994, Halliday and Matthiessen 2004) is the need for continuous contextualization: the study of the relationship between the strata (or *realization* in Halliday's terms, cf. Halliday and Matthiessen 2004:26), as instantiated in the form of text, has to consider the context, namely, the environment in which the text unfolds (Halliday and Hasan 1976). This constant need for contextualization undoubtedly recalls Firth's theory of meaning as function in context, which points out that the "progressive contextualization of linguistic facts" (Firth 1957b:36) is important and that "no study of meaning apart from a complete context can be taken seriously" in that the complete meaning of a word is always contextual (Firth 1935a:37). Nevertheless, although Firth (1957b:35, cf. Firth 1935b) often emphasizes the importance of "operating in context of situations which are typically recurrent, and repeatedly observable" and of placing such contexts of situation "in categories of some sort, sociological and linguistic, within the wider context of culture", as Halliday (1992a:24) clearly affirms, it is

Malinowski (1935)¹ who first pointed out that in order to understand a text, it is “necessary to extend the notion of “context” beyond the words and sentences on either side, and to include in it features of the non-linguistic environment”, which he labels the *context of situation* and the *context of culture* respectively (Halliday 1992a:24; cf. Section 1.1.3 for further details).

These two seminal ideas of the mutual dependence of the different levels of language and of the system and its environment, which have laid the basis for the Hallidayan systemic functional theory of language, have become two important aspects of what Stubbs (1996) calls the *Firth-Halliday-Sinclair line of development* (i.e. the framework within which the present research works). This line, which developed into corpus linguistics, operates within a contextual and functional framework which starts from Firth’s idea of *meaning* as being “subject to the general rule that each word when used in a new context is a new word” (Firth 1957b:190), and constantly highlights the interdependence of the different levels of language and of the consequent unpredictability of word meaning in isolation. As Sinclair says “words have many meanings, and there is no way of working out in advance which one is appropriate in a text” (Sinclair 2004a:137). The Sinclairian *idiom principle* and the Hoeyian idea of *priming* are examples that illustrate these principles at work.

1.1.1 The Idiom Principle

In order to explain Firth’s idea of continuous re-contextualization (Firth 1935b, 1957b:36) and the functional concept of the interdependence of lexis and grammar Sinclair (1991, 1996, 2004a:29) introduces the *idiom principle*, also known as the principle of *phraseological tendency* or *co-selection*, as opposed to the *open-choice principle* or the principle of *terminological tendency*. The former argues that “words enter into meaningful relations with other words around them” (Sinclair 2004a:25); the latter, that “words cannot remain perpetually independent in their patterning” (Sinclair 2004a:30; cf. also Hunston and Francis (2000) on the relationship between meaning and form). As Sinclair (2004a) puts it:

¹ According to Halliday (1992a:24), it is in Malinowski’s 1935 work that he introduces the notion of *context of situation* first, whereas, in fact, Malinowski introduces it before 1935, in his Supplement I, *The problem of meaning in primitive languages*, (cf. Malinowski 1927:306). In this work he maintains that the meaning of an expression “becomes only intelligible when it is placed within its *context of situation*” and he coins “an expression which indicates on the one hand that the conception of *context* has to be broadened and on the other that the *situation* in which words are uttered can never be passed over as irrelevant to the linguistic expression”.

Complete freedom of choice [...] of a single word is rare. So is complete determination. [...] I have called their linguistics correlates [...] the *open-choice principle* and the *idiom principle*. The preponderance of usage lies between the two. Some features of language patterning tend to favour one, some the other. Tending toward open choice is what we can dub the *terminological tendency*, which is the tendency for a word to have a fixed meaning in reference to the world, so that anyone wanting to name its referent would have little option but to use it, especially if the relationship works in both directions. Another tendency – almost the opposite – is the natural variation of the language, so that very little indeed can be regarded as fixed. Tending towards idiomativity is the *phraseological tendency*, where words tend to go together and make meanings by their combination. Here is collocation, and other features of idiomativity (Sinclair 2004a:29).

This systematic and expected co-occurrence of words, which recalls Firth's (1935b, 1957b) idea of language routine², and the new meanings made by such co-occurrences are demonstrated by lexico-grammatical features such as *collocations* (Firth 1957a, Leech 1974, Sinclair 1991, Hoey 1991, Stubbs 2001), *multi-word sequences* (Sinclair 1998, 2004a; Scott 1998, Biber *et al.* 1999, Hunston 2006, Stubbs 2006), *colligations* (Firth 1957a, Sinclair 2003, 2004a:174, Hoey 2005), *semantic prosody* and *semantic preference* (Louw 1993, Sinclair 2003, 2004a:174, Stubbs 2001, Partington 2004, Hunston 2007).

The notion of *collocation* is first introduced by Firth (1957a:14), who defines it as “actual words in habitual company”. Firth (1951a, 1957b) particularly emphasizes the habituality which distinguishes collocation and the limited possibility of co-occurrence of words, or, in Sinclairian terms, the *phraseological tendency* of language:

One of the meanings of *ass* is its habitual collocation with an immediately preceding *you silly*, and with other phrases of address or of personal reference. [...] There are only limited possibilities of collocation with preceding adjectives, among which the commonest are *silly, obstinate, stupid, awful*, occasionally *egregious* (Firth 1957b:195).

² Cf. (my emphasis): “We must take our facts from speech sequences, verbally complete in themselves and operating in contexts of situation which are **typical**, **recurrent**, and **repeatedly** observable”, Firth (1957b:35).

Firth's contextual approach to word meaning maintains that "meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words" (Firth 1957b:196). Later, other scholars give a slightly different definition of *collocation*: Leech (1974), for example, points out the psychological association "a word acquires on account of the meanings of words which tend to occur in its environment" (Leech 1974:20); Sinclair (1991:170), instead, emphasizes the textual trait of collocation, i.e. "the occurrence of two or more words within a short space of each other in a text"; and, both Hoey (1991) and Stubbs (2001) highlight its statistical aspect: i.e. the chance of relationship that "a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991:6-7), or, simply, "frequent co-occurrence" (Stubbs 2001:29). However, despite the different slants provided (i.e. contextual, psychological, textual, and statistical), what remains at the basis of the notion of collocation is the Firthian intuition that the meaning made by the co-occurrence of two items in a given context is a product of those two co-occurring words in that particular context, or in Hallidayan terms, "of the relationship between the system and its environment" (Halliday 2003b:196, cf. Halliday 1985c). An example of this creation of new meaning, provided by Sinclair (1998, 2004a:135), is the use of the adjective *white* which, when followed by the noun *wine*, implies a different color range from when it is used in isolation.

The notion of *multi-word sequences*, more technically called *lexical items* (Sinclair 1998, 2004a), *clusters* (Scott 1998, Scott and Tribble 2006), *lexical bundles* (Biber *et al.* 1999), *n-grams* (cf. Fletcher³), *sequences of words* (Hunston 2006), or *phrasal units* (Stubbs 2006), is directly linked to the Firthian notion of collocation in that it expands the category in terms of number of words, or *lexical items* (Sinclair 1998, 2004a) involved. Indeed, words do not only come in sets of 2 (as collocates do), but also in sets of 3, 4, or more, items and these items together create a meaning which is different from the meaning of the single items taken in isolation (Sinclair 2004a:134). Indeed, *multi-word sequences* such as *do you want to, I don't know what, I want to know, well that's what I* (cf. Biber 2006), for instance, are made up of words which, if taken in isolation, would have a different meaning and function from the whole cluster: these bundles, which are usually identified by frequency-driven approaches (which analyze, first of all, the most frequently recurring sequences of words in a text or

³ Cf. *PIE: Phrases in English*. On-line. Available from Internet, <http://pie.usna.edu>.

corpus), are generally incomplete grammatical structures which function as a unit in discourse, especially by bridging two structural units together (cf. Biber 2006: 133-135).

The third lexico-grammatical feature that illustrates the idiom principle at work is Firth's notion of *colligation*, which is a relation that words have at the grammatical level ("the inter-relation of grammatical categories in syntactical structure", Firth 1957a:15). Sinclair (2004a:174) and Hoey (2005:43) define it further: the former describes *colligation* as "the co-occurrence of words with grammatical choices", whereas the latter, deliberately recalling Firth⁴, as "the grammatical company a word or word sequence keeps (or avoids keeping)". Despite differences in wording, the above definitions stress the *phraseological tendency* of language in that they focus on restrictions on the grammatical choices accompanying a word. This restriction – called *relation* by Firth, *co-occurrence of choices* by Sinclair and *company* by Hoey – is illustrated by an example given by Sinclair (1996, 2004a:35), who claims that the lexical item *true feelings*, for instance, shows a strong (left) colligation with, or grammatical preference for, a possessive adjective, as in "... we try to communicate *our* true feelings to those around us..." and when it does not, it is still accompanied by another possessive construction such as *the true feelings of*.

Semantic prosody and *semantic preference*, the fourth and fifth lexico-grammatical features that exemplify the idiom principle, are usually attributed to Louw (1993) and Sinclair (2003, 2004a) respectively: the former term is defined as the "consistent aura of meaning with which a form is imbued by its collocates" (Louw 1993:158), and the latter as "the co-occurrence of words with semantic choices" (Sinclair 2004a:174). It is worth noting, however, that Louw (1993:158) himself attributes the notion of *semantic prosody* to Sinclair's (1987) intuition that items are habitually associated either with pleasant or unpleasant events⁵.

These two lexico-grammatical features interact: *semantic preference* contributes powerfully to building *semantic prosody*; and *semantic prosody* "dictates the general environment which constrains the preferential choice of the node item" (Partington

⁴ Cf. Firth's definition of collocation: "You shall know a word by the company it keeps!" (Firth 1957a:11).

⁵ Cf. Sinclair (1987:155-156) on the phrasal verb *set in* (my emphasis): "The most striking feature of this phrasal verb is the nature of the subjects. In general they refer to unpleasant states of affairs. Only three refer to the weather; a few are neutral, such as *reaction* and *trend*. The main vocabulary is *rot* (3), *decay*, *ill-will*, *decadence*, *impoverishment*, *infection*, *prejudice*, *vicious (circle)*, *rigor mortis*, *numbness*, *bitterness*, *mannerism*, *anticlimax*, *anarchy*, *disillusion*, *disillusionment*, *slump*. Not one of these is desirable or attractive".

2004:151). This can be illustrated by the *true feelings* example mentioned above. As Sinclair's (1996, 2004a:35) findings indicate, the expression *true feelings* usually displays a *semantic prosody* which is *negative*⁶, in that it tends to occur with expressions which suggest *reluctance* ("as in *will never reveal, prevent me from expressing, careful about expressing, less open about showing, guilty about expressing, etc.*", Sinclair 2004a:35) and *inability* ("as in *try to communicate, incapable of experiencing, unable to share*", Sinclair 2004a:35) and a *semantic preference* for verbs which are related to the semantics of *expression* (Sinclair 1996, 2004a:35). So, the *semantic preference* of the item *true feelings* for verbs which belong to the semantics of *expression* contributes powerfully to building its *semantic prosody*; in particular, since these verbs tend to express *reluctance* and *inability*, the *semantic prosody* of the item *true feelings* is negative. At the same time, the negative *semantic prosody* of the item *true feelings* dictates the general environment, i.e. the co-occurrence with the verbs which express *reluctance* and *inability*, which constrains its preferential choice for verbs which belong to the semantics of *expression*.

The terms *semantic prosody* and *semantic preference* have been further considered by Stubbs (2001), Partington (2004), and Hunston (2007), *inter alia*⁷, who, in line with Sinclair's discourse function of the unit of meaning (1991), all concur that it is necessary to take into account the discourse function of longer sequences, rather than merely the simple co-occurrence of two items. Stubbs (2001:111-12) describes *semantic* (or rather *discourse*) *prosody* as "a feature which extends over more than one unit in a linear string"; Partington (2004:132) underlines how its evaluative meaning spreads over "a unit of language which potentially goes well beyond the single orthographic word and is much less evident to the

⁶ Louw (1993) usually calls *positive* and *negative* prosodies *good* and *bad* respectively, whereas Partington (2004) also uses the pair *favourable* and *unfavourable*.

⁷ The notion of *semantic prosody* has also been criticized by Whitsitt (2005), for instance, who disagrees with the labels and analogies used by Louw (1993), such as the idea of semantic prosody extending from one context to another in the same way the vowels in the word *Amen* are imbued with a nasal quality because of their proximity to the nasals *m* and *n*: "He [referring to Louw 1993] claims that once the verb *set in* gets coloured with a negative meaning, it will not only always have that *colour*, but it will tend to only appear with words which have negative meanings, or be the word which "colours" other words with negative, *bad semantic prosody*. The analogy on which this argument is based, however, simply does not hold" (Whitsitt 2005:291). With this claim Whitsitt (2005:291) wishes to show that "the semantic prosodist is still faced with the problem of having to demonstrate on what grounds it can be claimed that a verb like *set in* is an empty form". It is worth noting, however, that one point that Louw (1993:158-159) makes is that the word *set in* can no longer be seen in isolation from its semantic prosody, and not that the color it acquires will last forever, indeed, one of the main milestones of corpus linguistics is that "a word which is used in a certain way in most contexts is not necessarily used in that way in all contexts" (Hunston 2007:252).

naked eye”, while Hunston (2007:266) suggests that the term semantic prosody “is best restricted to Sinclair’s use of it to refer to the discourse function of a unit of meaning”.

Terminology regarding the concepts of semantic prosody and preference is an area of contention: scholars disagree over the precise meanings of these terms and the extent to which evaluative attitude is involved. Stubbs (2001:111-12) prefers the label *discourse prosody*, rather than *semantic prosody*, pointing out that it often expresses “the speaker’s reason for making the utterance, and therefore identify[ies] functional discourse units” and uses *semantic preference* to express “relation between a lemma or word-form and a set of semantically related words”. Hunston (2007:266), instead, keeps the term *semantic prosody*, but suggests “that a different term, such as ‘semantic preference’ or perhaps ‘attitudinal preference’, should be used to refer to the frequent co-occurrence of a lexical item with items expressing a particular evaluative meaning”. Stubbs (2001:111-12) maintains that *semantic* (or *discourse*) *prosody* often expresses the speaker’s attitude (i.e. “the speaker’s reason for making the utterance”); Partington (2004:150) agrees that there is an evaluative or attitudinal slant to semantic prosodies, “used to express the speaker’s approval (good prosody) or disapproval (bad prosody) of whatever topic is momentarily the object of discourse” (cf. Sinclair 1996:87). On the other hand, Partington (2004:152) claims that *semantic prosody* is independent of individual speakers: in his view, this is because competent speakers of a language share the vast majority of lexical primings⁸ (cf. next Section), “otherwise communication would be impossible”.

1.1.2 Priming

The second concept which has grown out of the interdependence of both meaning and context and of lexis and grammar (cf. Section 2.1) is Hoey’s notion of *priming*, which explains the concepts mentioned so far from a slightly different, i.e. psychological, perspective. In Hoey’s words:

collocation is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random

⁸ I.e. the “psychological association between words” (Hoey 2005:5; cf. Section 1.1.2 on priming).

distribution [...]. We can only account for collocation if we assume that every word is mentally **primed** for collocational use. As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context. The same applies to word sequences built out of these words; these too become loaded with the contexts and co-texts in which they occur. I refer to this property as **nesting**, where the product of a priming becomes itself primed in ways that do not apply to the individual words making up the combination [...]. In this way, lexical items [...] and bundles [...] are created (Hoey 2005:5-11).

Hoey (2005) makes further hypotheses about lexical items; for instance, he suggests that they are not only collocationally primed, but they become primed also for semantic associations, colligation, and for textual position, and that they cannot be properly acquired unless they have all this priming. He also maintains that semantic association depends on the semantic set or class a lexical item occurs with, colligation on a grammar the lexical item tends to have, and textual position on the place (e.g. beginning of sentence, beginning of paragraph) a lexical item occurs in:

So, to illustrate, *result* is primed for collocation with *good*, it is primed for use as a noun or as a verb, it is primed for semantic association with positiveness (*a good result, a great result, an excellent result, a brilliant result, etc.*), and it is primed for use in certain grammatical contexts, e.g. definiteness (*the result v. a result*) [and] preliminary investigation suggests that, for example, *x years ago* has a powerful tendency to begin both paragraphs and texts⁹.

⁹Hoey, from the site <http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesofText.htm>, cf. also Hoey, 2005).

1.1.3 Implications of a Theory of Context

The lexico-grammatical features explained so far (i.e. collocations, multi-word sequences, colligations, semantic prosody and semantic preference), together with the psychological concept of priming, which have been identified by scholars from the Firth-Halliday-Sinclair tradition, have exemplified the idiom principle, or phraseological tendency of language. In particular, it has been shown that lexical items enter into meaningful relations with their environment: single words taken in isolation display different meanings and functions from the whole cluster, which also needs to be contextualized in order to be understood. It has also been demonstrated that lexical items have grammatical and semantic restrictions, or preferences.

The identification of this phraseological tendency in language is rather revolutionary in that it challenges the Saussurian idea of the linear (i.e. syntagmatic) nature of linguistic relations¹⁰ (Saussure 1972) and the Chomskyan notion that “grammar is autonomous and independent of meaning” (Chomsky 1957:17). It contests that the relation between semantics and syntax “can only be studied after the syntactic structure has been determined on independent grounds” (Chomsky 1957:17)¹¹. Indeed, the notions of collocations, multi-word sequences, colligations, semantic prosody and semantic preference not only demonstrate the interdependence of meaning and context, but also imply the consequent interconnection of the syntagmatic and paradigmatic axes (cf. Stubbs 1996, Sinclair 2004) and the existence of a lexico-grammatical interface (i.e. how semantics and syntax interface with each other).

Apart from being revolutionary, this perspective of the existence of an interaction between words and contexts and of the related need to contextualize linguistic facts can also be particularly useful: these lexico-grammatical associations like *collocates*, for example, can help to disambiguate differences between nearly equivalent grammatical structures or similar words. Biber, Conrad and Reppen’s (1998) analyses of *big* and *large* and *small* and *little* show

¹⁰ Cf. “In discourse [...] words acquire relations based on the linear nature of language because they are chained together” (Saussure 1972:123).

¹¹ In a personal email exchange with Prof. Chomsky dated September 21 2008, I had further (and more up-to-date) feedback on this point. Prof. Chomsky, indeed, backs up this claim highlighting that to him “grammatical status is independent of meaningfulness, as the examples illustrate (and innumerable others like them). It would follow, then, that the rules of grammar function independently of meaning - and there is massive evidence for that - although they provide the structures that determine the meaning of expressions”.

that considering the context¹² resets the possibilities of choice, since adjectives which are nearly synonymous in isolation, in fact, tend to co-occur with different words. This implies that lexico-grammatical associations help choose typical collocates such as *big toe* and *large number*, for instance, rather than unusual collocates such as *big number* or *large toe*.

A further example of the fact that this limited collocational choice, and lexicogrammatical associations in general, are particularly useful for disambiguating differences is provided by the notion of translation-in-context. Halliday himself (1992a:15) insists that linguistics can offer a theory of context, in that it can be useful to translation in terms of what is possible, rather than in terms of rules that should be applied (Halliday 1992a:15). Taking the context into account resets the probabilities; consequently, choices that may be less likely in isolation may be preferred in larger contexts (Halliday 1992a:17). Indeed, if collocation acts as a constraint, then it can “help justify the restriction of the field” (Firth 1957:180) by resetting the probabilities and choices of translation (Halliday 1992a:17).

To explain this concept further, Halliday (1992a) provides an example regarding the equivalence of morphemes across languages: usually, the most probable Italian equivalent of the morpheme *-ly* at the end of an English word is *-mente* (cf. Halliday 1992a:17), but in order to decide whether this equivalent holds, the suffix *-ly* must be seen within the context of an English word. “For instance, in *likely* the *-ly* at the end is not rendered by *-mente*: *likely* is *probabile* not *probabilmente*. So the most immediate context, that of the next rank in the grammar, has given us the information we need to make another choice” (Halliday 1992a:17). This example is extremely relevant in that it demonstrates that knowing the associations between words and grammatical structures can provide relevant clues both to understanding and translating language. In particular, it clarifies the idea that in order to grasp the meaning

¹² That is: *little*, similarly to *big*, usually co-occurs with concrete, often animate, nouns, whereas *small*, similarly to *large*, with nouns indicating quantity. *Small* has a stronger association with predicative position than *little*, and this association is especially strong in conversation; however, predicative *small* in conversation is often used to characterize physical size, like *little* in attributive position, but with different functions: with predicative *small*, the main point of the utterance is to identify “smallness” as an important feature of the noun being described; conversely, *little* in attributive position provides an identifying characteristic of the noun being described, but the utterance itself has another purpose (Biber, Conrad and Reppen 1998). The following examples are given by Biber, Conrad and Reppen (1998:94) to compare *small* in predicative position (examples A and B) to *little* in attributive position (examples C and E):

- A. She's *small* and really skinny
- B. He's really *small*, isn't he?
- C. She's known me since I've been a *little* girl.
- D. Well, he's like any *little* kid I think.

of even a single morpheme, the context as well as the different levels of language need to be taken into account (e.g. the grammatical level provides evidence that *likely* is an adjective and not an adverb like most other words ending in *-ly*). Of course, when translating, it is not enough merely to take account of the mutual dependence of the different levels of language, because other non-linguistic features, such as the *context of situation* and the *context of culture*¹³ (Halliday 1992:24), play an important role: the more precisely one can define the context of situation, the more exactly one may predict the properties of a text in that situation, especially if one includes the context of culture (Halliday and Hasan 1976:22-23). Indeed, as the following quote by Malinowski (1927:301-302) illustrates, it is arduous to understand a text written in a language by people living outside that language community, even if it is translated into their language, due to the fact that each message brings more meanings than those expressed through the words. These can only be understood by considering the environment of the text, defined in the value systems and ideology of that specific culture:

Instead of translating, of inserting simply an English word for a native one, we are faced by a long and not altogether simple process of describing wide fields of custom, of special psychology and of tribal organization which correspond to one term or another. We see that linguistic analysis inevitably leads us into the study of all subjects covered by Ethnographic field-work (Malinowski 1927: 301-2).

The application of this functional approach, which considers translation as a process in context (Halliday 1992a:15), is illustrated by a simple but effective example in Ulrych (1992), who points out that:

¹³ As explained by Halliday and Hasan (1976:21) (cf. also Section 1.1), the concept of context of situation was, first, formulated by Malinowski (1927) and, then, elaborated by Firth (1950, 1957b) and it “refers to all those extra-linguistic factors which have some bearing on the text itself”. More specifically, in Hallidayan terms, the *context of situation* is made up of the *field*, the *mode* and the *tenor*: the field being “the total event, in which the text is functioning, together with the purposive activity of the speaker or writer”; the mode being “the function of the text in the event, including therefore both the channel taken by the language – spoken or written, extempore or prepared – and its genre, or rhetorical mode, as narrative, didactic, persuasive, ‘phatic communion’ and so on”; the tenor being “the type of role interaction, the set of relevant social relations, permanent and temporary, among the participants involved” (Halliday and Hasan 1976:22).

the Italian *ciao* is used as an informal greeting equally on arrival and departure. Thus, in translating *ciao* into English, translators need first to analyse the SL [Source Language] text to establish whether the context is one of coming or one of going. From this they can deduce whether the function is saying hello or saying goodbye (Ulrych 1992:69).

It can, then, be concluded that, since lexical items enter into meaningful relations both with their linguistic and extra-linguistic environment, the non-linear (i.e. exclusively syntagmatic) nature of linguistic relations and the autonomy of grammar and meaning are excluded from the present functional framework. Conversely, to grasp the functional meaning of lexical items, the various interdependent levels of language must be taken into account, together with their extra-linguistic context. In order to do so, the present work is based on a data-oriented description of language which has the advantage of accessing computational tools, like software and corpora, (cf. next Sections) and which takes account of the environment (both linguistic and extra-linguistic, when necessary) in order to discover the functional meaning of the items investigated.

1.2 Authentic Data and Corpora

The following quotation from Sinclair introduces and also summarizes the content of this section, which advocates a data-oriented description of language through computer analysis, namely, a functional descriptive approach that investigates meaning as function in context through empirical data offered by corpora:

In summary I am advocating that we should trust the text. We should be open to what it may tell us. We should not impose our ideas on it, except perhaps just to get started. Until we see what the preliminary results are, we should apply only frameworks that are loose and flexible, in order to accommodate the new information that will come from the text. We should expect to encounter unusual phenomena; we should accept that a large part of our linguistic behaviour is subliminal, and that therefore we may find a lot of surprises. We should search for models that are especially appropriate to the study of texts and discourse.

The study of language is moving into an era in which the exploitation of modern computers will be at the centre of progress. The machines can be harnessed in order to test our hypotheses, they can show us things that we may not already know and even things which shake our faith quite a bit in established models, and which may cause us to revise our ideas very substantially. In all of this my plea is to trust the text. (Sinclair 2004a:23).

In particular, the present work investigates language “in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences” (Stubbs 1996:28, cf. also Halliday 1992b, 2003c:208; Biber, Conrad, and Reppen 1998; McEnery and Gabrielatos 2006), in terms of probability of occurrence (cf. Kennedy 1998:270, Halliday 1993), and through both qualitative and quantitative use of corpora (cf. McEnery and Wilson 1996; Biber, Conrad, and Reppen 1998; Sinclair 2006). Section 1.2.1 illustrates the reasons for choosing authentic-data oriented analyses by comparing them to those based on intuition (Section 1.2.1.1), by pointing out the advantages of qualitative and quantitative analyses (Section 1.2.1.2), empirical data (Section 1.2.1.3) and electronic processing (Section 1.2.1.4). Section 1.2.2, instead, highlights the drawbacks linked to this approach and possible solutions to them.

1.2.1 Advantages of Authentic-Data Oriented Analyses

The reason for opting for an authentic-data oriented analysis is strictly linked to the following factors: authentic data allow for descriptions of naturally occurring combinations of words as opposed to the limiting and sometimes deviating traits of intuition; they offer both quantitative and qualitative analyses; they are empirical and lead toward descriptivism as opposed to prescriptivism; they can be collected in large databases (i.e. corpora) in electronic format and can, consequently, be easily and quickly processed, replicated, and shared.

1.2.1.1 Authentic Data vs. Intuition

As pointed out by Svartvik (2007:16) (cf. also Stubbs 1996 and Sinclair 2004a, *inter alia*), “detailed analysis of a corpus consisting of real-life language” is “very much swimming against

the tide of the mainstream Chomskian view of language”¹⁴; Chomsky (1965), indeed, argues that intuition and isolated sentences¹⁵ are the basis of linguistics¹⁶. Conversely, in the view presented here, words are not considered to “have fixed meanings which are recorded, once and for all, in dictionaries” (Stubbs 2001:13); rather, they acquire them, which sometimes involves modification of meaning, “according to the social and linguistic contexts in which they are used” (Stubbs 2001:13).

One of the main advantages of analyzing data is that it brings about the description of words in context and, especially, reflects the actual combinatorial possibilities of language, which introspection cannot discern (cf. Sinclair 2006): in Johansson’s (2007:51) words, it is “a way of seeing” (cf. also Stubbs 2001:66). Conversely, data collected through introspection may be blind and deviate from naturally occurring data (Börjas 2006:11), especially for two reasons: firstly, the problem with introspection (conceived both as *informant testing*, or *evidence of secondary sources*, such as reports by speakers on their usage, and as *introspection by the linguist*, as Sinclair 1991:39 classifies them) is that it does not give evidence about language usage; on the contrary, it only offers broad ideas about it (Sinclair 1991:39). Secondly, as pointed out by Biber, Conrad, and Reppen (1998), finding patterns of use and analyzing contextual factors can present methodological difficulties because one is more likely to be looking for typical patterns than unusual occurrences.

Furthermore, analyses of large amounts of language, rather than conclusions based on a few speakers’ idiosyncrasies, are usually suggested (Stubbs 2001; Sinclair 2004a, 2004b) and intuition cannot either keep track of or reproduce every kind of occurrence (Stubbs 2001). Another strong point of empirical data stored in computerized corpora is that they can be easily processed, mined, and reproduced by software, consequently providing a solution to this problem (cf. Section 1.2.1.4).

¹⁴ Cf. Chomsky quoted in Aarts (2001:6): [corpus linguistics] “doesn’t exist”, “you don’t take a corpus, you ask questions. [...] Otherwise you just get junk. [...] You want an answer to a non-trivial question, you’ve got to go beyond looking at data”.

¹⁵ Chomsky (1957:17) also maintains that “grammar is autonomous and independent of meaning”, as pointed out above.

¹⁶ In a personal email exchange dated September 21 2008 which I had with Prof. Chomsky, he maintains that “It’s a special case of the observation that “data science” doesn’t exist. Data certainly exist, and they are of interest for understanding insofar as they are selected to respond to queries. That’s how all rational inquiry works. That’s why scientists do experiments -- asking questions of nature. It’s why paleoanthropologists look for certain configurations of terrain and artefacts and bones, not others. Etc. It’s universal. How could linguistics be different? Corpus linguistics also seeks answers to very specific questions, and discards data that are irrelevant. That is, it’s not corpus linguistics”.

All this shows that conclusions based on intuition or on anecdotal evidence can, therefore, be unreliable and limiting (cf. also Ulrych 1999a:76 on the importance of recognizing both “the limitations of intuition” and those of “an exclusively experience-based approach”). Nevertheless, it is worth emphasizing that, even though native intuition as well as traditional concepts and categories are downplayed in a corpus-driven analysis, data alone cannot suffice. This is true for two reasons: first of all, analyses are bound to the linguist’s subjectivity in that, as Biber, Conrad, and Reppen (1998:4) maintain, it is “the human analyst” who has to “make difficult linguistic judgments”, consequently, total objectivity is rarely achieved (Stubbs 2001; Sinclair 2006); second, functional (qualitative) interpretations are also an essential step in any corpus study because data need to be explained (Biber, Conrad, and Reppen 1998:4; Sinclair 2006; cf. Section 1.2.1.2 on qualitative and quantitative analyses). This implies that intuition cannot be totally abandoned in linguistic analyses; rather, it necessarily plays a crucial role in them (Hoffmann 2004, Sinclair 2006, Johansson 2007), although it does not provide new evidence. Corpora, then, become “resource[s] against which to test intuitions and motor[s] which can help to generate them” (Partington 1998:1), or in Leech’s (1991:74) words, corpus use becomes “a question of corpus *plus* intuition, rather than of corpus *or* intuition”.

1.2.1.2 Authentic Data, Quantitative and Qualitative Analyses

As pointed out by Halliday (2003a:23), quantitative features are “an inherent part of the meaning potential of language”; however, it is essential to go beyond quantitative analyses, which computerized corpora can provide, by including functional (qualitative) interpretations of quantitative data which explain them (Biber, Conrad, and Reppen 1998:5-9; Aarts 2001). The former are relevant because, involving data counting and classification, they allow for statistical reliability and generalization, whereas the latter allow for detailed descriptions and involve an in-depth analysis and understanding of the features analyzed (McEnery and Wilson 1996; Sinclair 1996; Biber, Conrad, and Reppen 1998; Kennedy 1998). The advantage of integrating quantitative findings and functional, qualitative interpretations and descriptions derives from the fact that qualitative and quantitative analyses acquire more strength by being complementary, as corpus linguists such as McEnery and Wilson (1996), Biber, Conrad, and Reppen (1998), Kennedy (1998), Aarts (2001), and

Sinclair (2006), for example, all maintain.

1.2.1.3 Authentic Data as Empirical Data

As first set out by Francis (1993) and then established by Tognini-Bonelli (2001), the corpus approach can be either corpus-based or corpus-driven. The “term corpus-based is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study” (Tognini-Bonelli 2001:65). The corpus, therefore, offers the chance of a quantitative extension to a linguistic theory, but the theory itself remains uninfluenced by the data observed (Tognini-Bonelli 2001:70). The corpus-driven approach, conversely, is an empirical, holistic approach, which constructs the theory step by step in the presence of the evidence (Tognini-Bonelli 2001:17). In other words, within the corpus-based approach the theoretical background pre-exists the corpus examination and is not questioned (cf. Tognini-Bonelli 2001:71; Sinclair 2006), whereas within the corpus-driven approach the theory totally depends on the evidence (Tognini-Bonelli 2001:84), which may give unexpected answers and change ideas and descriptions which were formulated before the analysis of the data (Sinclair 2006); in Johansson’s (2007:55) words, it offers a way of making “new discoveries”, (cf. Higgins’ 1991 *serendipity principle*¹⁷).

The main advantage of the corpus-driven approach, which is the guiding principle of the present analysis, is the potential it has to shed empirical light on research. This, following Renouf (2007), may be labeled *the corpus linguist’s scientific driver*, namely, “the desire to undertake an empirically-based methodological cycle, beginning with curiosity based on introspection, intuition and probably data observation” (Renouf 2007:29) and confirms Halliday’s view that systemic-functional linguistic principles and practices corresponding “fairly closely to the sort of things that scientists do” (Halliday 2003c:200).

Another strong point in favor of empiricism is that it is also accompanied by the value of the exhaustiveness, representativeness, explicitness, and replicability of corpus data. Exhaustiveness is envisaged in the examination of all the data retrieved (Sinclair 2006);

¹⁷ According to the *serendipity principle*, new discoveries may emerge accidentally while looking for something else entirely, which may lead to new hypotheses. The term originates from the British writer Horace Walpole and from a Persian fairy-tale about three princesses from Serendip (Arabian name for Sri Lanka), who made discoveries they did not expect to make (Christoffersen 2004).

representativeness is manifested by the fact that “a corpus seeks to represent a language or some part of a language. [...] The representativeness of the corpus [...] determines the kinds of research questions that can be addressed and the generalizability of the results of the research. For example, a corpus composed primarily of news reportage would not allow a general investigation of variation in English” (Biber, Conrad, and Reppen 1998: 246). Explicitness is linked to a high level of detail given, that is to say that “with reference to the retrieval of data from a corpus, the input, the output and the interpretation should be set out in detail” (Sinclair 2006). Replicability is scientifically important, being the traditional basis of the empirical method. Making data publicly available is also fundamental, as Sinclair (2006:15) points out: “it is vital to distinguish between evidence about language which is shared and that which is personal” (cf. also Wynne 2004).

1.2.1.4 Authentic Data and Electronic Processing

As stated above, analyses of large amounts of language are usually time-consuming and difficult to monitor. The use of empirical data stored in computerized corpora, which can be easily and quickly processed by computerized software (Hoffmann 2004; Mahlberg 2006) “provided that the computer is properly instructed” (Hoffmann 2004:190), offers a solution to this problem. So, first and foremost, the most advantage of corpus linguistics is that raw data can be turned into a database and be processed (Stubbs 2001). This implies:

- . time-saving procedures and wide-ranging storing: the speed of technological development of computers (such as the ever-increasing size of memory and speed of access, the specificity of some linguistic software programs, etc.) nowadays allows for wide-ranging storing, which is of particular importance in that it enables one to check a large number of occurrences; as Sinclair (2004a) explains, the bigger the corpus, the better it is, even though size is not a warrant of representativeness;

- . fast and complex analyses: another advantage of the technological development of computers is that they provide fast, complex analyses and reliable calculations, which could not be possible, or would be extremely time-consuming, without calculators and software programs (Börjas 2006, Mahlberg 2006);

- . replication of experimental results: another strong point of empirical data stored in computerized corpora is that they allow for replication of experimental results, which is “an

essential procedure for checking and refining knowledge” (Stubbs 2001:123) in that replication guarantees more objectivity, precision and exhaustiveness (Sinclair 1991, 2004a; Stubbs 2001; Wynne 2004; Börjas 2006), all scientifically important features of empirical research. Conversely, “if language study is based on introspective data from the individual linguist, a genuinely reproducible experiment is rarely possible, since neither data nor methods are independent of the analyst” (Stubbs 2001:123);

. reusability and sharing of resources: the use of computerized data is also advantageous because the electronic format can be re-used and moved without data loss. This ensures interchange and reusability of resources within the scientific community (Stubbs 2001).

Finally, as illustrated in the previous Sections (i.e. Sections 1.2.1.1, 1.2.1.2 and 1.2.1.3), although it is worth underlining that “computer-assisted methods of text analysis cannot interpret texts for us” (Stubbs 2001:124), they can provide empirical, quantitative and qualitative investigations of naturally occurring language which intuition and introspection alone cannot offer, or, in Stubbs’ (2001:124) words, “they can provide, for subsequent human interpretation, new kinds of evidence”.

1.2.2 Corpora Drawbacks and Possible Solutions

There are, of course, drawbacks in corpus studies: to quote Sinclair (2006:19), “there are many potential pitfalls along the road”. Indeed, first of all, the corpus cannot capture the whole language system, in that, as Halliday (2003a:25) puts it (cf. also Renouf 1997), “we are far from being able to measure the size of language in any meaningful way. All we can say is that a language is a vast, open-ended system of meaning potential, constantly renewing itself in interaction with its eco-social environment”; consequently any corpus will surely lack some language constructions. Second, the software may be prone to error, the variables may spoil the data, the linguist’s subjectivity is difficult to eliminate, and there is the risk of “reinventing the wheel” (Sinclair 2006:9).

To begin with the first drawback mentioned, it is widely acknowledged that no corpus, no matter how large or how carefully designed, can have exactly the same characteristics as the language itself (Sinclair 2004c), a corpus cannot capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact, there are no

such things as *correct proportions* of components of an unlimited population (Sinclair 2004c). The only solution to this is to be aware of it, and to be mindful of language potentials (cf. Halliday 2003a:25). Especially, one has to be conscious of the fact that the absence of a particular construction, for instance, does not imply that that construction is absent from the specific language analyzed or the whole language itself: “as a probe becomes more delicate and complex, it is likely to retrieve fewer and fewer instances, though often fascinating ones, and it is tempting to continue beyond the point where the evidence justifies the finding. It is tempting, too, sometimes to point out the absence of things that might reasonably be expected, and occasionally justified, but corpora are still far too small for absence to be anything more than a hint” (Sinclair 2006:19). A possible way around this drawback is to increase the amount of the data investigated, either by enlarging the corpus or by checking that particular construction in other corpora and redoing the analyses (Börjas 2006:11, Sinclair 2006).

However, it must be acknowledged that collecting spoken data is relatively difficult, for the following reasons. First, representative speakers who agree to be recorded have to be found; second, they have to be recorded in such a way that their recording can then be easily accessed and heard; finally, these recordings need to be transcribed so as to be investigated with corpus linguistic software.

Another concern is that “sometimes a little-known quirk of the software produces results that are not what they seem” (Sinclair 2006:19). Moreover, results may be spoilt by the linguist’s subjectivity: “in all these problem areas, the researcher is responsible for the validity of the statements made. Using an established corpus in a straightforward manner and ensuring that there is reasonable numerical support for claims is a fairly safe stratagem, but it is fair to say that corpora are in danger of being exploited way beyond their ability to deliver reliable results” (Sinclair 2006:19). A feasible solution to these two problems is given by replicability. Indeed, a key scientific process to check whether results are consistent is to repeat the analyses (Sinclair 2006).

A further issue for corpus studies is the possible presence of uncontrollable variables, which may diminish the value of the results. The possible solution is “to cut the variables down” (Sinclair 2006:4), even though this may undoubtedly be a hard task if there are too many variables to take into account.

Finally, the unpredictability of the corpus-driven approach may lead to the risk of

“reinventing the wheel”, i.e. of spending effort and time to come to the conclusion that the previous descriptions were correct after all. However, as Sinclair (2006:9-10) puts it, “this kind of verification is a normal part of scientific method in general; more importantly, however, if corpus-driven research is able to devise categories of description which fit corpus patterns more neatly and comprehensively than pre-corpus categories, then these descriptive categories will constitute a strong argument for making modifications to theories in order to align theory, description and data in a more direct and informative way than they are placed at the moment”.

1.3 Units of Analysis and Tools

One of the very first decisions which determined the object of the present research was to define the unit(s) of the analysis, i.e. the “observation(s)” for the study (cf. also Sinclair’s units of description; Sinclair 1998, 2004a:148). In corpus investigations, the unit of analysis is typically one of two kinds: either a single text (if the goal of the research is to describe a type of text from a group of texts) or the occurrences of a linguistic feature. In the first case, each observation is a text, whereas in the second, each observation is an occurrence of the structure in question (Biber, Conrad, and Reppen 1998:269). To measure the observation and to carry out a number of quantitative and qualitative analyses, it is necessary to code a large sample of constructions and to consider each occurrence as a separate observation (Biber, Conrad, and Reppen 1998:269). In order to do so, a corpus and a software program are required. The former contains, for instance, the occurrences of the linguistic feature under examination and consequent information on its frequency and pragmatic functions, the latter helps to retrieve them.

In this dissertation the units of the analysis are of both kinds: the text, or rather the texts, i.e. American face-to-face and movie conversation, and the linguistic feature, i.e. discourse marker *you know*. First, the two conversational domains are investigated at a macro-level through Multi-Dimensional analyses (multi- in that they involve more than one dimension, cf. Section 1.3.1); then, the occurrences of the discourse marker *you know* are investigated in the two conversational domains mentioned at a micro-level through mono-analyses (mono- in that they involve only one item, cf. Section 1.3.1). These analyses include both the linguistic feature itself and the two conversational domains in which it occurs, in

that describing the pragmatic functions of *you know* does not only elucidate its nature, but also the nature of the conversational domains where it occurs. The data for American face-to-face conversation were retrieved from the *Longman Spoken American Corpus* (cf. Section 1.3.2), whereas the data for American movie conversation come from the *American Movie Corpus* (cf. Section 1.3.3). As for the approach adopted, corpus-driven criteria are followed, apart from when the occurrences of the items analyzed are too numerous. Under this circumstance, the analyses are performed on a sample selection of the data and/or via hypothesis testing, following the suggestions of Sinclair (1999) and Hunston (2002:52):

Sinclair (1999) advocates selecting 30 random lines, and noting the patterns in them, then selecting a different 30, noting the new patterns, then another 30 and so on, until further selections of 30 lines no longer yield anything new. An adaptation of this method is ‘hypothesis testing’, in which a small selection of lines is used as a basis for a set of hypotheses about patterns. Other searches are then employed to test those hypotheses and form new ones.

The *Biber grammatical tagger*, the *SAS software package*, the software programs *MonoConc Pro Version 2.0* (published by Athelstan) and *Oxford Wordsmith Tools 4.0* (developed by Scott 1988, cf. also *Oxford Wordsmith Tools 4.0* guide) were also used for information retrieval (cf. Section 1.3.1).

1.3.1 Multi-Dimensional and Micro Analyses

Multi-Dimensional (often MD in quotes) analysis is an approach developed by Biber (1988) which applies multivariate statistical techniques¹⁸ to determine co-occurrence relations among linguistic features and thus investigate register¹⁹ variation. Biber’s (1988:63-64) claim

¹⁸ Multivariate statistics involves observation and analysis of more than one statistical variable at a time; cf. Izenman (2008:17): “Multivariate data consist of multiple measurements, observations, or responses obtained on a collection of selected variables. The types of variables usually encountered often depend upon those who collect the data (the domain experts), possibly together with some statistical colleagues; for it is these people who actively decide which variables are of interest in studying a particular phenomenon. In other circumstances, data are collected automatically and routinely without a research direction in mind, using software that records every observation or transaction made regardless of whether it may be important or not”.

¹⁹ The way in which the term *register* is used in the present research follows Biber (1995) and Biber and Conrad

is based on the assumption that frequently co-occurring linguistic features in texts share at least one communicative function, and that it is possible to identify a unified dimension underlying each set of co-occurring linguistic features:

In factor analysis, a large number of original variables, in this case the frequencies of linguistic features, are reduced to a small set of derived variables, the ‘factors’. [...] Each factor represents an area of high shared variance in the data, a grouping of linguistic features that co-occur with a frequency. The factors are linear combinations of the original variables, derived from a correlation matrix of all variables (Biber 1988:79).

The Multi-Dimensional approach, in other words, via factor analysis, reduces “a large number of linguistic variables to a few basic parameters of linguistic variation” (Biber 2004:19) and identifies the co-occurrence patterns among specific linguistic features called *Dimensions* (Biber 1988; Biber, Conrad and Reppen 1998):

In MD analysis the co-occurrence patterns among a large number of linguistic features are identified with the statistical technique known as factor analysis. In a factor analysis, the correlations among a large number of variables (i.e., the linguistic features) are identified, and the variables that are distributed in similar ways are grouped together. Each group of variables is a factor – which is then interpreted functionally as a “dimension” of variation (Biber, Conrad and Reppen 1998:278).

The *Dimensions* considered in the present research are represented by the following

(2001): it is close to the term *genre* and covers situationally defined varieties (cf. Biber 1995:7 and Biber and Conrad 2001:3). That is, register is defined by situational characteristics (cf. Halliday and Hasan 1976:21) like “differences in purpose, interactiveness, production circumstances, relations among participants, etc.” (Biber 1995:7). Conversely, the term *text type* refers here to text categories defined in strictly linguistic terms regardless of any non-linguistic factors like purpose, topic, or interactiveness, even though, after being identified on formal grounds, they can be interpreted functionally (cf. Biber 1995:10). Consequently, “text types are defined such that the texts within each type are maximally similar with respect to their linguistic characteristics (lexical, morphological, and syntactic), while the types are maximally distinct with respect to their linguistic characteristics” (Biber 1995:10).

five *Factors*²⁰ (Biber 1988):

. Factor 1 represents a dimension labeled “Informational versus Involved Production”, that is to say, a dimension which marks “high informational density and exact informational content versus affective, interactional, and generalized content” (Biber 1988:107). Factor 1 involves two parameters: the primary purpose of the writer/speaker, which can be either informational or interactive, affective, and involved; and the production circumstances, which can be characterized by either careful editing, precision in lexical choices and an integrated textual structure, or by generalized lexical choices and fragmented presentation of information.

. Factor 2 represents a dimension labeled “Narrative versus Non-narrative Concerns”, that is to say, a dimension which “can be considered as distinguishing narrative discourse from other type of discourse” (Biber 1988:109). *Narrative concerns* are marked by the presence of past time, third person animate referents, reported speech, and details, whereas *non-narrative concerns* are marked by immediate time and attributive nominal elaboration.

. Factor 3 represents a dimension labeled “Explicit versus Situation-Dependent Reference”, that is to say, a dimension which distinguishes “between highly explicit, context-independent reference and nonspecific, situation-dependent reference” (Biber 1988:110). *Wh relative clauses*, for instance, specify the identity referents explicitly, whereas *time* and *place adverbials* are dependent on referential inferences (Biber 1988:110).

. Factor 4 represents a dimension labeled “Overt Expression of Persuasion”, that is to say, a dimension which “marks the degree to which persuasion is marked overtly” (Biber 1988:111). Biber holds that prediction, necessity, possibility modals, together with infinitives, conditional subordination, suasive verbs, and split auxiliaries mark persuasion.

. Factor 5 represents a dimension labeled “Abstract versus Non-abstract Information”, that is to say, a dimension which “seems to mark informational discourse that is abstract, technical, and formal versus other types of discourse” (Biber 1988:113). The use of conjuncts, agentless passive verbs, by-passives, passive postnominal modifiers, *inter alia*, have positive weights on this factor.

²⁰ Biber (1988 and 1995) and Conrad and Biber (2001) consider also Factor 6 and 7, namely, the dimensions about “On-line Informational Elaboration Marking Stance” and “Academic Hedging”. They have not been taken into account here for they are considered still tentative by the literature because they are too difficult to interpret (cf. Conrad and Biber 2001:39). It is worth noting, however, that face-to-face conversation is usually unmarked in the use of the features associated with these dimensions.

The parameters just illustrated are considered *Dimensions* in that they define “continuums of variation rather than discrete poles” (Biber 1988:9). This means that Multi-Dimensional analysis describes texts that are to be interpreted as more or less formal, narrative, explicit, etc. rather than either formal or non-formal, narrative or non-narrative, explicit or situation-dependent, etc. This is clarified by the following texts from Biber (1988:10-12), which show that, even though text 1 (i.e. an example of *conversation*) and text 2 (i.e. an example of *scientific exposition*) seem to offer dichotomies (i.e. conversation is common, unplanned and interactive, whereas scientific exposition is specialized, planned and non-interactive), by looking at text 3 (an example of *panel discussion*) these parameters define continuum dimensions. Consequently, text 1 can be described as less specialized, planned and more interactive than text 2, and text 3 as being between the two.

Text 1. Conversation (Biber 1988:10)

<i>A:</i>	<i>I had a bottle of ordinary Courage's light ale, which I always used to like, and still don't dislike, at Simon Hale's the other day – simply because I'm, mm, going through a lean period at the moment waiting for this next five gallons to be ready, you know.</i>	1 2 3 4 5 6
<i>B:</i>	<i>mm</i>	7
<i>A:</i>	<i>It's just in the bottle stage. You saw it the other night.</i>	8
<i>B:</i>	<i>yeah</i>	9
<i>A:</i>	<i>and, mm I mean, when you get used to that beer, which at its best is simply, you know, superb, it really is.</i>	10 11
<i>B:</i>	<i>mm</i>	12
<i>A:</i>	<i>you know, I've really got it now, really, you know, got it to a T.</i>	13 14
<i>B:</i>	<i>yeah</i>	15
<i>A:</i>	<i>and mm, oh, there's no, there's no comparison. It tasted so watery, you know, lifeless.</i>	16 17
<i>B:</i>	<i>mm</i>	18

Text 2. Scientific exposition (Biber 1988:10)

<i>Evidence has been presented for a supposed randomness in the movement of plankton animals. If valid, this implies that migrations involve kineses rather than taxes (Chapter 10). However, the data cited in support of this idea comprise without exception observations made in the laboratory.</i>	1 2 3 4 5
--	-----------------------

Text 3. Panel discussion (Biber 1988:11-12)

W: *But Mr. Nabarro, we know that you believe this.*

L: *quite*

W: *The strange fact is, that you still haven't given us a reason for it.*

The only reason you've given for us is, if I may spell it out to you once more, is the following:

the only crime for which this punishment was a punishment, after its abolition, decreased for eleven years.

You base on this the inference that if it had been applied to crimes it never had been applied to, they wouldn't have increased.

Now this seems to me totally tortuous.

In order to apply the Multi-Dimensional approach, the following eight methodological steps need to be considered (Biber 1988, 1995, 2004):

1. The corpus design, collection, and transcription (in the case of spoken texts) and input into the computer;
2. Identification of the linguistic features and of their functional associations to be included in the analysis;
3. Development of computer software programs which tag all relevant linguistic features in the corpus;
4. The automatic tagging of the corpus and editing of the texts to check whether the linguistic features are accurately identified;
5. Counting of each linguistic feature in each text of the corpus via additional computer programs;
6. Factor analysis of the co-occurrence patterns among linguistic features;
7. Functional interpretation of the factors as underlying dimensions of variation;
8. Computing of the dimension scores for each text; comparison of the mean dimension scores for each register to analyze the salient linguistic similarities and differences among the registers being studied.

The importance of Biber's (1988) analysis depends on a number of advantages. First and foremost, it is reliable: apart from Biber's own work (Biber 1988; Biber 1995, Biber 2006), a large number of experiments have been carried out to evaluate it (cf. Atkinson 2001;

Biber and Finegan 2001a, 2001b; Conrad 2001; Helt 2001; Reppen 2001a; Rey 2001; Quaglio 2004) and have shown that, even when split corpora are investigated, factor analysis provides nearly the same dimensions of variation, as long as the samples of the corpora include an equivalent range of register variation (Biber 2004:16). Biber and Finegan (2001a) and Atkinson (2001), for example, have respectively focused on the historical evolution of register by analyzing diachronic relations among speech-based and written registers, and scientific discourse across history, whereas Biber (1988 and 2006) and Reppen (2001a) have analyzed register variation in speech and writing, and Biber (1995) has provided a cross-linguistic comparison of register variation. Furthermore, Biber and Finegan (2001b) and Conrad (2001) have investigated specialized domains such as medical research articles (Biber and Finegan 2001b) and textbooks and journal articles in biology and history (Conrad 2001). Helt (2001), Rey (2001) and Quaglio (2004), instead, have studied dialect variation by comparing British and American spoken English (Helt 2001), male and female language in the American television series *Star Trek* (Rey (2001) and the language of the TV series *Friends* to face-to-face conversation (Quaglio 2004)

Secondly, Biber's (1988) Multi-Dimensional approach is also important in that, via computer programs, it can predict the extent to which two linguistic features vary when they occur together:

A large negative correlation indicates that two features co-vary in a systematic, complementary fashion, i.e. the presence of the one is highly associated with the absence of the other. A large positive correlation indicates that the two features systematically occur together (Biber 1988:79).

This means that if the factor analysis of a corpus reveals, for example, that the occurrence of first person pronouns in a text is high, it can then be expected that questions will occur to a similar extent; conversely, when first person pronouns are absent from a text, it is likely that questions are absent too (Biber 1988:80). Interestingly, previous research has also led to the hypothesis of the existence of universal dimensions of register variation, in that some dimensions seem to occur across languages and across general and restricted discourse domains (Biber 2004:17).

Last but not least, through factor analysis, Biber's (1988) Multi-Dimensional approach provides, quantitative methods which empirically confirm the notion of co-occurrence by identifying and interpreting patterns which co-occur as underlying *dimensions* of variation:

In the interpretation of a factor, and underlying functional dimension is sought to explain the co-occurrence pattern among features identified by the factor. That is, it is claimed that a cluster of features co-occur frequently in texts because they are serving some common function in those texts (Biber 1988:91).

In the present research, two distinct analyses are provided: at a macro and more generic level, Multi-Dimensional factor analyses are presented as a way of determining the text type²¹ which movie language belongs to, and comparing it to face-to-face conversation; at a micro level, instead, the occurrences of the discourse marker *you know* are checked in two spoken corpora. In both cases, quantitative and qualitative techniques are used: in the Multi-Dimensional analyses, the co-occurrence patterns as underlying dimensions of variation are identified, first, quantitatively via factor analyses, then, qualitatively via functional interpretation. In the analyses of *you know*, the frequency of its occurrence are calculated first and, then, quantitative and qualitative considerations on the pragmatic function that this discourse marker acquires in the two conversation domains considered are illustrated. More specifically, the quantitative analyses of *you know* focus on the number of occurrences within and without the utterance position (i.e. the specific frequency according to initial, medial, final position and the general frequency have been calculated), whereas the qualitative, pragmatic analyses focus on the kind of functions it acquires in these contexts by considering its context and the lexical and functional items with which it occurs.

As regards information retrieval, the texts were kindly processed for Multi-Dimensional analyses by Douglas Biber with the tagger he developed (i.e. the *Biber grammatical tagger*) and the *SAS software package* for statistical analyses he adapted for linguistic studies. The *Biber grammatical tagger* was used to identify grammatical features to be processed by the *SAS software package*; this subsequently turned them into the underlying

²¹ Cf. Note 19.

Dimensions characterizing the two conversational domains investigated here.

The software programs *MonoConc Pro Version 2.0* (published by Athelstan) and *Oxford Wordsmith Tools 4.0* (developed by Scott 1988, cf. also *Oxford Wordsmith Tools 4.0* guide), instead, were used to investigate the occurrences of *you know*. Both programs offer similar features, such as the ability to generate wordlists, concordances and collocations, and they both can handle large tagged or untagged corpora. The reason for using two software programs which can perform similar tasks was to compensate for their individual limits: *MonoConc Pro*, for instance, can split the screen display and expand the context of the node by highlighting the line in a more user-friendly way than *Wordsmith Tools 4.0*, while *Wordsmith Tools 4.0* provides useful plots which give information about the distribution of an occurrence in a single text or across texts, and cluster information (cf. Reppen 2001b on the differences between the two software programs). In particular, *WordList*, *Concord*, and *Plot* were used from *Wordsmith Tools 4.0*: the first created lists of all the words or word-clusters in the texts, set out in alphabetical or frequency order; the second retrieved words or phrases in context to see the company they keep; and the third provided information about the distribution of an occurrence in a single text or across texts (Scott 1998, Scott and Tribble 2006).

1.3.2 The Longman Spoken American Corpus

The corpus of American face-to-face conversation, i.e. the *Longman Spoken American Corpus* (henceforth LSAC), used for the present analyses to represent spontaneous conversation is taken from the *Longman Spoken and Written English Corpus*²². It belongs, together with the *Longman Written American Corpus*, to the *Longman Corpus Network*. In particular, the *Longman Spoken American Corpus*, which is the five-million-word-corpus used for the present research, is owned by Pearson Education and was gathered by Professor Jack Du Bois and his team at the University of California at Santa Barbara (UCSB): at least four hours of the daily conversations of American speakers from all regions of the US, chosen as representative for gender, age, ethnicity, and education, were recorded as unobtrusively as possible by project workers with tape recorders. The conversations took place over periods of at least four days.

²² I was kindly given access to the corpus by Prof. Douglas Biber and Prof. Randi Reppen during my visit as a guest scholar at Northern Arizona University in April-May 2008.

The tapes were subsequently edited to eliminate silences and garbled material and then transcribed. So as to guarantee anonymity, names, addresses, and phone numbers that were mentioned during the recordings were not transcribed, even though records of the situations being recorded, and of the details of the participants were kept (Stern 2005).

1.3.3 The American Movie Corpus

The **American Movie Corpus** (henceforth AMC) is a corpus I specifically developed for the study of American movie language. In technical terms, it is a *sample parallel bilingual corpus*: *sample*, in that it does not claim to be representative of the whole variety under examination, i.e. movie language, but rather aims to provide a representative snapshot of it²³. It is *parallel* and *bilingual* because it is made up of original texts (i.e. original American movies) plus their translated versions (i.e. the relative dubbed Italian movies)²⁴.

There were two reasons for building up a movie corpus: first, in spite of the relatively large amount of available spoken American English corpora (cf. the Bank of English, MICASE, Santa Barbara Corpus, etc.), to my knowledge, no corpus provided appropriate material for American movie language analysis; second, the scripts²⁵ which are easily accessible and freely downloadable from the web turned out to be inappropriate for this kind of investigation, in that their transcriptions of speech differ considerably from what is actually said in the movies. To give an example, the total amount of words transcribed for the movie *Shallow Hal* is 11,490, whereas the script retrieved from the web²⁶ contains 10,660 words. In particular, there are 49 occurrences of *you know* and 31 of *I mean* in the transcription, whereas they occur respectively 38 and 23 times in the web script. The following extracts, which are from the AMC and from the web, show the extent of this difference²⁷: Extract 1 demonstrates that the same scene has the same content in the two transcriptions, but different wording; Extract 2, instead, shows that the movie starts in a completely different

²³ Cf. <http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>.

²⁴ For a discussion of types of corpora see cf. Ulrych (1999b:64-65).

²⁵ They are also called *screenplay* or *transcripts*.

²⁶ Cf. http://www.script-o-rama.com/movie_scripts/s/shallow-hal-script-transcript-paltrow.html

²⁷ Furthermore, it is worth noting that the scriptwriter him/herself sometimes puts a note underlying that the web script has not been written to represent the actual words in the movie: "This transcript is not trying to get the movie word for word, but close to it. This transcript is for reading purposes only!"; source: <http://www.awesomofilm.com/script/MI2.html>

way from the script.

Extract 1. A scene from *The Devil Wears Prada*: same content, different wording

Extract from the AMC		Extract from the Screenplay by Peter Hedges ²⁸
Andrea	Hi. Uh, I have an appointment with Emily Charlton?	<p>EXT. RECEPTION -- LATER</p> <p>ANDY is trying to arrange herself on the uncomfortable sofa when suddenly a taller, thinner and, amazingly, more groomed version of the women in the room walks in.</p>
Emily	Andrea Sachs?	<p>This is EMILY, who looks the part of the sleek fashionista, but is propelled by a core of barely tamped down anxiety.</p> <p>EMILY Um... Andrea Barnes?</p>
Andrea	Yes.	<p>EMILY looks up. Their eyes meet. As EMILY takes in how different ANDY looks from everyone else...</p> <p>...ANDY springs up and follows her down the hallway.</p>
Emily	Great. Human Resources certainly has an odd sense of humor. Follow me.	<p>INT. RUNWAY -- DAY</p> <p>EMILY walks ANDY down the hall.</p> <p>EMILY Who put you up for this job?</p> <p>ANDY Human Resources sent me.</p> <p>EMILY They do have an odd sense of humor.</p>
Emily	Okay, so I was Miranda's second assistant... but her first assistant recently got promoted, and so now I'm the first.	<p>At the end of a long corridor is a bullpen -- two desks outside a large corner office.</p> <p>ANDY can only see part of the corner office, but it is seductively bright, sending light streaming into the bullpen.</p>
Andrea	Oh, and you're replacing yourself.	<p>ANDY and EMILY sit down.</p> <p>EMILY (cont'd) Miranda has two assistants -- I'm the first, and we're interviewing for the second, junior assistant. (pauses, dramatic) Miranda is an amazing woman, a legend. (MORE)</p>
Emily	Well, I am trying. Miranda sacked the last two girls after only a few weeks. We need to find someone who can survive here. Do you understand?	
Andrea	Yeah. Of course. Who's Miranda?	
Emily	Oh, my God. I will pretend you did not just ask me that. She's the editor in chief of Runway, not to mention a legend. You work a year for her, and you can get a job at any magazine you want. A million girls would kill for this job.	
Andrea	It sounds like a great opportunity. I'd love to be considered.	
Emily	Andrea, Runway is a fashion magazine so an interest in fashion is crucial.	

²⁸ Source: www.dailyscript.com/scripts/devil_wears_prada.pdf

		<p>EMILY (cont'd) Working for her sets you up to work anywhere in publishing. A million girls would kill for this job.</p> <p>ANDY Sounds great.</p> <p>EMILY The thing is, Andy, we are a fashion magazine and an interest in fashion is crucial.</p> <p>ANDY What makes you think I'm not interested in fashion?</p> <p>EMILY gives her a look. Suddenly, EMILY'S Blackberry goes off. She gasps.</p> <p>EMILY Oh my God. No. No, no, no.</p> <p>ANDY What's wrong?</p>
Andrea	What makes you think I'm not interested in fashion?	

Extract 2. The very first words uttered in *Catwoman*: totally different content

	Extract from the AMC	Extract from the Screenplay by Dan Waters²⁹
Patience	<p>It all started on the day that I died. If there had been an obituary, it would have described the unremarkable life of an unremarkable woman, survived by no-one. But there was no obituary, because the day that I died was also the day I started to live. But that comes later....</p>	<p>EDNA Yes?</p> <p>PATIENCE Hi, Edna Powers? (off her nod) I'm Patience Price, I called about adopting a cat? I saw your flyer at my vet's office --</p> <p>EDNA Oh yes, do come inside.</p>

In front of this evidence, the decision was taken to manually transcribe the original and dubbed Italian versions of a number of American movies (cf. also Figure 2). A 204,636-word corpus³⁰ was built up (i.e. nearly 22 hours of movie speech in both American English and Italian) according to the criteria outlined in Section 1.3.3.1. The AMC consists of the following transcribed movies: *Mission: Impossible II*, or *M:I2* (John Woo 2000); *Erin Brockovich* (Steven Soderbergh 2000); *Me, Myself & Irene* (Bobby and Peter Farrelly 2000); *Meet the Parents* (Jay Roach 2000); *Finding Forrester* (Gus Van Sant 2000); *Shallow Hal* (Bobby and Peter Farrelly 2001); *Ocean's Eleven* (Steven Soderbergh 2001); *One Hour Photo*

²⁹ Source: <http://www.dailyscript.com/scripts/catwoman.pdf>

³⁰ More specifically the original (i.e. English) component is made up of 104,530 words, whereas the dubbed (i.e. Italian) one consists of 100,106 words.

(Mark Romanek 2002); *The Matrix Reloaded* (Andy and Larry Wachowsky 2003); *Catwoman* (Pitof 2004); *The Devil Wears Prada* (David Frankel 2006). The present research, which does not consider dubbing, is based on the original versions of the movies, namely, on the 104,530-word component.

Figure 2. Movies of the American Movie Corpus



1.3.3.1 Corpus Building Criteria

In principle, any collection of texts can be called a corpus (corpus being Latin for *body*), hence a corpus is any body of text (cf. McEnery and Wilson 1996); however, in corpus linguistics terms a corpus is not meant to be a mere collection of texts, but precise compilation criteria exist (Sinclair 1991, McEnery and Wilson 1996, Renouf 1997, cf. also Sections 1.2 and 1.2.2). For the compilation of the AMC, the four main characteristics of a modern corpus, as suggested by Sinclair (1991), McEnery and Wilson (1996), Kennedy (1998) were taken into account: sampling and representativeness (and consequent balance), standard reference, finite size, and machine-readable format.

As pointed out above (cf. Section 1.2.2), one of the limits of corpus studies is that no corpus, regardless of its size or design, can precisely reflect and capture the language as a whole and accurately represent it (cf. Sinclair 2004b, Renouf 1997, Kennedy 1998). It is

nevertheless feasible to build up a sample corpus which provides a representative snapshot of realistic data, providing an acceptable view of the tendencies of the language population one wishes to study by limiting the population itself (cf. Biber 1993). The notions of representativeness and balance are, in the final analysis, matters of judgment and can only be approximate”; indeed, “generalizations are an essential part of science” (Kennedy 1998:62).

It can also be argued that every movie script could be suitable for the study of movie language, in that in order to become a movie, a script must be performable as such. Consequently, every movie script that has turned into a movie can be said to represent movie language. In practice, however, one has to deal with the problem of variables (cf. Section 1.2.2), in that in order to have controllable and comparable data, variables need to be limited. As a consequence, since the interest of the present research is the investigation of specific features of contemporary American movie conversation compared to American face-to-face conversation, the movies selected had to satisfy certain parameters. To be selected for the AMC, movies had to:

- (a) be produced in the United States from 2000 on;
- (b) be acted/spoken mostly in American English;
- (c) not be set in previous centuries and eras;
- (d) have ordinary life settings.

Parameters (a) and (b) determine the kind of domain and variety under examination, i.e. American movie language. Parameter (b) also reflects the idea of dialog in action, that is to say, dialog had to be present in the movie selected (e.g. narrated movies, documentaries and, of course, mute movies were excluded). Parameter (a), together with (c), also implies the contemporaneity of movies and parameter (c) also guarantees that the language spoken in the movies selected is the ordinary language of ordinary people: specialized language is not the focus of the present study, consequently, movies based on political debates, academic speeches, legal language and other specific domains were not included. It is worth noting that, even though some of the characters of the movies selected have extraordinary powers – e.g. Neo in *The Matrix Reloaded* –, they are people who lead ordinary lives – Neo, for instance, works in information technology.

Finally, movies were categorized as belonging to genres³¹ of comedy, non-comedy, or border-line (when the categorization could not be clear-cut). This rather simplistic categorization was introduced to represent a wider spectrum of movie language, first to satisfy the balance and representativeness parameters, which require the full range of linguistic variation existing in the language (Biber 1993, Kennedy 1998) and, second, to see whether genre variation influences the frequency of the spoken devices analyzed. In particular, considering that movie genre is difficult to define (cf. Table 1 which shows how different sources categorize movies in different ways) especially due to the fact that movies usually do not belong to only one genre, the AMC components were defined along a comedy/non-comedy continuum which took into account the suggestions given by Morandini (2007) and by the Internet Movie Database³² (henceforth IMDB). Table 1 illustrates the AMC components grouped according to their genre: four movies are considered to be 100% comedies, and indeed both Morandini (2007) and the IMDB classify them as such; for the same reason, 4 others are considered to be 100% non-comedies; and 3 have been inserted as not genre specific, being characterized in different ways by the two classifications selected³³ (i.e. Morandini 2007 and the IMDB).

Table 1. AMC movies and genres

MOVIE	IMDB	Morandini (2007)	CONTINUUM
Shallow Hal	Comedy / Drama / Romance	Comedy	100% COMEDY
Meet the Parents	Comedy	Comedy	
Me, Myself & Irene	Comedy	Comic	
The Devil Wears Prada	Comedy / Drama	Comedy / Drama	
Ocean's Eleven	Comedy	Thriller	50% COMEDY 50% NON- COMEDY
Erin Brockovich	Biography / Drama	Comedy	
Finding Forrester	Drama	Comedy / Drama	
One Hour Photo	Drama / Mystery / Thriller	Drama	
Catwoman	Action / Crime / Fantasy	Sci-Fi	100% NON-COMEDY
Mission: Impossible II	Action / Adventure / Thriller	Adventure	
The Matrix Reloaded	Action / Sci-Fi / Thriller	Sci-Fi	

The idea behind building the movie corpus is to set the foundation stones of a large, standard reference corpus for future studies of the movie language it represents (cf. Section

³¹ Cf. Note 19.

³² <http://www.imdb.com/>

³³ Due to their non-clear-cut status, these movies have been labeled borderline movies in the analysis sections.

1.3.3.2 below). However, often the finite number of words is determined at the beginning of a corpus-building project (Sinclair 1991, McEnery and Wilson 1996), so a finite number of movies were selected.

Eleven movies were transcribed, and this depended exclusively on the time given to the transcription process: in the three-year-research-project, the first year was dedicated to the corpus building so that there would be time to gather data within the corpus, study them, and draw relative conclusions. To quote Kennedy (1998:81), “as is known by anyone who has had to transcribe substantial amounts of text, transcription can be time-consuming, expensive and fraught with difficulties”

As stated above (cf. Section 1.3.3), the transcriptions of the eleven American movies, together with their dubbed Italian versions, make up a 204,636-word corpus (nearly 44 hours of movie conversation). By some standards, the AMC is a small corpus. Sinclair (2004a:189) holds that big corpora are usually favored for linguistic research due to the fact that they have a better chance of including regularities of language. However, as Sinclair (2004a) himself, Biber (1993), and Kennedy (1991:68) point out, “a huge corpus does not necessarily ‘represent’ a language or a variety of a language any better than a smaller corpus”; indeed, everything depends on the patterns present in the corpus and the consequent generalizations that can be made about them (e.g. for the study of prosody, for example, “a corpus of 100,000 words will usually be enough to make generalizations for most descriptive purposes”, Kennedy 1991:68).

Another issue relevant to the size of the AMC is comparability: having the same number of words in the two corpora selected would certainly have been more ideal; however, in the time available, it was not possible to manually transcribe the same amount of words as in the Longman corpus. Nevertheless, despite the different size of the two corpora considered, data comparison was feasible by means of normalizing/*norming* the numbers (Biber, Conrad, and Reppen 1998:263). Indeed, as Biber, Conrad, and Reppen (1998:263) report, it is always possible to adjust data via normalization when corpora are not of the same length and frequency counts are not directly comparable. Obviously, the total number of words in each text was considered when frequency counts were normed. Specifically, the raw frequency count was divided by the number of words in the text, and then multiplied by 1.000 words, which was the basis chosen for norming.

The last factor taken into account in compiling the corpus is linked to its format: the

manually transcribed movie dialogs are stored in *.txt*, *.doc*, and *.xls* electronic files. This choice of machine-readable format derives from the reasons pointed out in Sections 1.2.1.2, 1.2.1.3, and 1.2.1.4. Machine-readable corpora favor both quantitative and qualitative data studies, allow for the storing of a wide range of data that can be searched and manipulated at speed and easily enriched with extra information. Furthermore, such corpora allow for objectivity and replicability of the studies, and ensure that information can be exchanged within the scientific community, and the data re-used. The choice of the *.txt*, *.doc*, and *.xls* formats depended on computing factors such as data processing and information retrieving. In particular, the *.txt* files were necessary for the *WordSmith Tools 4.0* software (i.e. for concordances and frequency counting), whereas the *.doc* and the *.xls* files were compiled to so as to include extra information such as the name of speakers, setting and relevant extra-linguistic features, which were not included in the *WordSmith Tools 4.0* analyses (cf. also Section 1.2.1).

1.3.3.2 Standardization and Transcription Criteria

As stated above, one of the main advantages of computerized data is that they provide replicability, which means that successive studies will not need to re-computerize the information. However, to ensure possible interchange of information and reusability of the resources within the scientific community, corpus-building has to follow standard data-storing and representativeness criteria. Standardization, indeed, is another fundamental component of corpus studies (Johansson 1993).

For the AMC annotation, a verbatim record of what is actually said in the movies was written, which is generally called orthographic transcription. In practice, the movies were watched and the dialogs were carefully transcribed. Then, for the sake of accuracy, they were double-checked by native speakers of English and Italian who were not involved in the transcription. The reason for choosing orthographic transcription was three-fold. First of all, it provides a representation of spoken language which is simple to read and understand (compared to IPA transcription, for instance, orthographic transcription is easier because it requires less effort and knowledge, cf. Halliday 1985b); secondly, it allows for immediate computing processes such as frequency and concordancing information retrieval (orthographic transcription, indeed, is the format which concordancers usually read); thirdly,

even though orthographic transcription is “an imperfect written approximation of a speech event” (Kennedy 1998:82), and cannot capture all the features of spoken conversation (Halliday 1985b, Wichmann 2007), it forms the basis for all other transcriptions and annotations. This means that the corpus can easily be enriched with extra information any time that future research requires it. This also reflects Sinclair’s clean-text policy (2004b:83, 1989) “to keep the text as it is, unprocessed and clean of any other codes”: in corpus-driven linguistics you do not use pre-tagged text, but you process the raw text directly so that the actual patterns of the uncontaminated text, ie. the actual text units, can then be observed and manipulated (Sinclair 2004a:191).

In view of the importance of standardization and consistency for the interchange and reusability of resources (Johansson 1993), the AMC orthographic transcription followed some of the international standards for transcription given by the Linguistic Data Consortium³⁴ (henceforth LDC) that are usually used with spoken corpora (cf. the Santa Barbara Corpus of Standard American English, for instance). So as to provide a standard format for data interchange and maintain compatibility with existing standards (Johansson 1993), the AMC transcription was based on the speaker identification and orthographic transcription conventions summarized in Table 2, 3 and 4³⁵:

³⁴ "The Linguistic Data Consortium is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes. The University of Pennsylvania is the LDC's host institution. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by grant IRI-9528587 from the Information and Intelligent Systems division of the National Science Foundation" (<http://www.ldc.upenn.edu/About/>).

³⁵ Cf. also LDC site: <http://projects.ldc.upenn.edu/SBCSAE/transcription/csae-conventions.html#ortho>

Table 2. Speaker identification and orthographic transcription conventions used for the AMC compilation

Speaker Identification	At the beginning of each transcript, the speaker is given a unique identifier if the name is not present. In this case the speaker's gender is also indicated.
Capitalization	Capitalization is used as an aid for human comprehension of the text. The accepted standard way to capitalize words, including words at the beginning of a sentence, proper names, and so on are followed.
Abbreviations	When abbreviations are used as part of a personal title, they can remain as abbreviations: <i>Mr. Brown</i> <i>Mrs. Jones</i> <i>Dr. Spock</i> However, when they are used in any other context, they are written out in full, e.g.: <i>I went to the junior league game.</i> <i>I'm going home to see the missus</i> <i>I went to the doctor, and all he said was, don't worry, it's natural.</i> <i>Hey mister, do you know how to get to the stadium?</i>
Contractions and Apostrophe -s	Table 3 below illustrates what is considered standard written English with respect to contractions.

Table 3. Transcription conventions used for the AMC compilation
for contractions and apostrophe s³⁶

Complete words	Contraction allowed
I have	I've
Cannot	can't
will not	won't
you have	you've
could not	couldn't
we will	we'll
should have	should've
it is	it's
Marvin - possessive	Marvin's
going to	Gonna
want to	Wanna
she is	she's
Marvin is	Marvin's
Marvin has	Marvin's

³⁶ The contractions in bold are not allowed in the LDC transcription guides. However, they were kept in the present research for two main reasons: firstly because they reflect what is actually said in the movies; secondly because they are also present in the *Longman Spoken American Corpus*, which is the corpus used for the present comparative study.

CHAPTER 2. FACE-TO-FACE CONVERSATION

The present chapter deals with face-to-face, or spontaneous, conversation and aims to illustrate some of its typical traits, focusing especially on discourse markers, in particular on *you know*. The intent is to offer a theoretical overview based on corpus studies. More specifically, Section 2.1 points out the reasons why spoken language has not been studied until recently and provides a taxonomy of the spoken domain. Section 2.2 describes the key features of spontaneous conversation outlining a spectrum of determinants that characterize it. Section 2.3 focuses on a typical phenomenon of conversation, i.e. discourse markers. In particular, Section 2.3.1 illustrates the problems related to their terminology, classification, and approaches, whereas Section 2.3.2 highlights the general traits that discourse markers commonly display. Section 2.3.3, instead, focuses specifically on the discourse marker *you know*, by investigating its functional categorization, providing a new, simpler classification, and describing its functions in combination with the turn position. The chapter concludes with a summary of the present framework (Section 2.4).

2.1 Spoken and Written Language

Although “a central tenet of twentieth-century linguistics” is that spoken language has priority over written language (Miller and Weinert 1998:4), that its primacy has to be perceived in terms of its occurrence in human societies (McCarthy 2003:15-16) – i.e. it is “the most commonplace, everyday variety of language” (Biber *et al.* 1999:1038) –, and that “the spoken language corpus is a primary resource for enabling us to theorize about the lexicogrammatical stratum in language – and thereby about language as a whole” (Halliday 2005:158), the grammar of conversation has been little researched until recently. The main contributions to spoken language have derived especially from scholars working with the field of pragmatics, who, focusing on “meaning-in-interaction” (Quaglio and Biber 2006:692), have provided insights into speech acts (Austin 1962, Searle 1969), implicature (Grice 1975), politeness (Brown and Levinson 1987), *inter alia*; and others working on Conversational Analysis who, dealing with *talk-in-interaction* (Thomas 1995, Quaglio and Biber 2006), have described patterns such as turn-takings (Sacks 1992; Ford, Fox, and Thompson 2002) and adjacency pairs (Sacks 1992), for instance.

Such limited and unbalanced interest was due to a number of reasons. First, as the Greek origin of the word *grammar* itself shows, the Western grammatical tradition is founded almost exclusively on the study of written language (grammar meaning *a letter*, i.e. *a piece of writing*, or *written mark* – cf. Biber *et al.* 1999 and OED³⁷); consequently, there has been no real choice, throughout most of the history of linguistics: “to study text, as data, meant studying written text; and written text had to serve as the window, not just into written language but into language” (Halliday 2005:159). Second, such research was not feasible before the event of sizeable computer corpora (Biber *et al.* 1999:1038; Miller 2006:670); or rather, before the event of the tape recorder, linguists had no means of capturing spoken language (Halliday 2005):

[...] to accumulate enough spoken language in a form in which it could be managed in very large quantities, we needed a second great technical innovation, the computer; but in celebrating the computerized corpus we should not forget that it was the tape recorder that broke through the sound barrier (the barrier to arresting speech sound, that is) and made the enterprise of spoken language research possible (Halliday 2005:157-158).

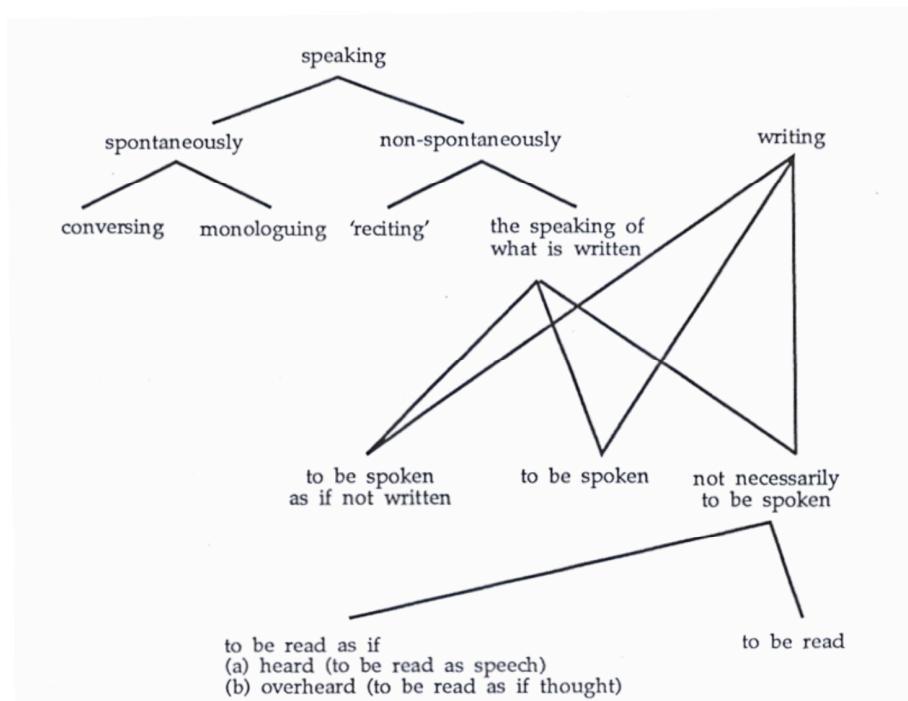
Third, even with the event of sizeable computer corpora, spoken language is more difficult to observe and to codify (McCarthy 2003:15-16); indeed, even if one transcribes spoken language, “one is always dealing with an imperfect product, especially compared to the accuracy with which the latest optical text scanners can quickly gobble up vast amounts of written text and deposit them in machine-readable form” (McCarthy 2003:13; cf. Halliday 2005:162).

Before proceeding, it is worth underlining that there are many types of speaking. The first difference (cf. Figure 3 from Gregory and Carroll 1978, cf. also Taylor 1999), deals with the spontaneity of the domain, which leads to further sub-categorization of spoken language: speaking *spontaneously*, indeed, may imply either *conversing* or *monologuing*, whereas speaking *non-spontaneously* may imply either *reciting* or *speaking of what is to be spoken as if not written*, of what is *to be spoken*, or of what is *not to be spoken*, which are usually based on *writing*. The kind of conversation illustrated in this chapter is spontaneous conversation, which, in Gregory and

³⁷ Oxford English Dictionary, available online at: <http://0-dictionary.oed.com.millennium.unicatt.it/>

Carroll's (1978) terms, is the kind of *speaking* which implies *conversing spontaneously* with an interlocutor without relying on any kind of writing or planning.

Figure 3. Spoken language sub-categories (Gregory and Carroll 1978:47)



2.2 Key Features of Spontaneous Conversation

Biber *et al.* (1999:1041) maintain that it is rather difficult to characterize conversation in terms of communicative goals or social functions; indeed, “the most that can be claimed is that it is a pervasive activity among human beings, and that its primary function appears to be to establish and maintain social cohesion through the sharing of experience, although secondarily it may promote other goals such as entertainment (e.g. through jokes and narratives), exchange information and control of others’ behaviour.” However, in spite of this lack and of the fact that “we are still very much in the age of exploration” (Chafe 1982:35), a spectrum of “determinants of conversation” (Biber *et al.* 1999:1041) can be identified so as to determine some of the features that characterize spontaneous conversation:

a. **Spontaneous conversation takes place in the spoken medium** in that it comes about through the oral-auditory channel (Biber *et al.* 1999:1041), and consequently involves supra-segmental features like pitch, rhythm, and voice quality (Miller 2006:673). At

the same time it also occurs with **non-verbal paralinguistic features** such as gestures, body postures, facial expressions, and eye-gaze/-contact which still signal information (Erman 1987:3; Miller and Weinert 1998:22; Bercelli 1999:97; Contento 1999:269; Miller 2006:673);

b. Spontaneous conversation takes place in real time, impromptu, with no opportunity for editing (Chafe 1982; Biber *et al.* 1999:1048; Miller and Weinert 1998:22; McCarthy 2003:109; Quaglio and Biber 2006:702; Miller 2006:672); indeed, speech is evanescent and ephemeral in its nature (Taylor 1999, Cameron 2001): “it consists of sound waves in the air, and sound begins to fade away as soon as it is produced” (Cameron 2001:31); consequently, if somebody utters *hello Rory*, for instance, by the time (s)he gets to *Rory* it is no longer possible to hear *hello*. The hearer must therefore process the utterance as it happens, in *real time* (Cameron 2001:31). Besides, it is subject to the limitations of short-term memory in both speaker and hearer (Miller and Weinert 1998:22; Miller 2006:673);

c. Spontaneous conversation usually takes place in a shared context in that the conversation participants usually share a large amount of contextual background, including specific social, cultural, and institutional knowledge; consequently, it does not need elaboration of meaning (Biber *et al.* 1999:1042; Bercelli 1999:96; Quaglio and Biber 2006:705);

d. Spontaneous conversation is interactive, continuous, and expressive of politeness, emotion, and attitude in that it is co-constructed by at least two interlocutors, i.e. a speaker and a hearer, who dynamically shape their expression to the ongoing exchange. This to-and-fro movement of conversation between the interlocutors is especially evident in utterance-response sequences, or adjacency pairs, which ”may be either symmetric, as in the case of one greeting echoing another, or asymmetric, such as a sequence of question followed by answer” (Biber *et al.* 1999:1045); in the routine use of discourse markers and similar devices such as interjections, response forms and vocatives, *inter alia*, which signal the dynamic and interactive role of the speaker’s utterance (Biber *et al.* 1999:1046; cf. also Bazzanella 1990 and Gavioli 1999); and in the use of “polite or respectful language in exchanges such as requests, greetings, offers, and apologies” (Biber *et al.* 1999:1047; cf. also Stame 1999, Brown and Levinson 1987). The continuity of conversation, instead, can be seen in the use of pragmatic expressions like gap fillers, discourse markers, hedges, tags, back channels, connectors, and interjections, *inter alia* (Erman 1987, Biber *et al.*

1999, Taylor 1999, Aijmer and Stenström 2005, Quaglio and Biber 2006, Redeker 2006).

The presence of supra-segmental and paralinguistic features and shared knowledge, i.e. features (a) and (c), imply reliance on reference and implicit non-elaborated meaning. Indeed, they are features which still signal information and allow conversation to be marked by simplified grammatical structures and a reduction of the number of words uttered (e.g. the use of non-causal or grammatically fragmentary components such as ‘stand-alone’ words which “rely heavily for their interpretation on situational factors”; Biber *et al.* 1999:1042), a high frequency of reference (e.g. pronouns, as contrasted with a low frequency of nouns) which can be supplied by mutual, or shared, knowledge (Biber *et al.* 1999:1042).

In much the same way, the absence of grammatical elaboration (such as pre-modification and post-modification and lexical density, *inter alia* – Halliday 1985, Taylor 1999, Biber *et al.* 1999) and referential specificity (as compared to the large use of vagueness and hedging strategies, for example) also arises from the reliance on context (Biber *et al.* 1999:1045) and on the face-to-face mode interaction (Chafe 1982:45; Miller and Weinert 1998:22; Miller 2006:673); indeed ”in drawing heavily on implicit meaning, conversation forgoes the need for the lexical and syntactic elaboration commonly found in written expository registers” (Biber *et al.* 1999:1044).

The on-the-fly trait of spontaneous conversation, i.e. feature (b), typically gives way to what has been called *normal dysfluency* (Biber *et al.* 1999:1048) and *fragmented language* (Chafe 1982:39). As Biber *et al.* (1999:1048) point out, “it is quite natural for a speaker’s flow to be impaired by pauses, hesitators (*er, um*), and repetitions such as *I – I – I* at points where the need to keep talking [...] threatens to run ahead of mental planning, and the planning needs to catch up [...]” (cf. also McCarthy 2003:112).

Undoubtedly, even though people do not “have time to mold a succession of ideas into a more complex, coherent, integrated whole” (Chafe 1982:37) when speaking, some degree of planning may be involved and “speed of communication can vary a great deal according to the needs of encoding and decoding” (Biber *et al.* 1999:1048). This, for example, happens when the speaker knows what to say or when the speaker and hearer share knowledge: “planning runs ahead of speech production” (Biber *et al.* 1999:1048). However, if on the one hand written language is “fostered by the greater amount of time available” (Chafe 1982:45), on the other, “the faster space of spoken language” (Chafe 1982:45) constrains speakers to reduce the length of what they have to say to save time and energy: “speed of

repartee, making an opportune remark, getting ‘a word in edgeways’ in a lively dialogue, or reaching the point quickly, may all add urgency to the spoken word” (Biber *et al.* 1999:1048).

Finally, taking place in real time and being interactive, i.e. features (b) and (d), lead speakers to tend to repeat the same repertoire of expressions relying on “stereotyped, prefabricated sequences of words” (i.e. lexical bundles, cf. Chapter 1) (Biber *et al.* 1999:1049; cf. also Tannen 1982; Bazzanella 1999; McCarthy 2003; Halliday 2005). “Time pressure makes it more difficult for speakers to exploit the full innovative power of grammar and lexicon” (Biber *et al.* 1999:1049) and repetition can help them to *buy time* to plan the next chunk (Cameron 2001:34) relying on “well-worn, prefabricated word sequences, readily accessible from memory” (Biber *et al.* 1999:1049). For the same reason, spontaneous speech typically contains redundant information (Cameron 2001:34). In much the same way, speakers employ a large number of connectors, gap fillers, hedges, tags, back channels, interjections, discourse markers, etc., to keep the conversation going (Erman 1987, Biber *et al.* 1999).

Studies on register variation (Chafe 1982; Halliday 1985; Biber and Finegan 1986; Biber 1988; Miller and Weinert 1998; McCarthy 1999; Halliday 2005; Biber 2006) have shown that spoken and written language differ in their use of nouns and verbs, especially because noun phrases are much more complex in written than in spoken texts and because very many of the words of spoken language “clearly belong to the traditional province of grammar/function words, in that they are devoid of lexical content” (McCarthy 1999:5; cf. also Halliday 1985 and Halliday 2005 on *lexical density*). In particular, Biber and Finegan (1986), via the use of multivariate statistical techniques such as factor and cluster analyses³⁸ (Biber 1985), have identified face-to-face conversation with what they call *Cluster 1*, namely, the highly interactive, situated and immediate text type. In other words, face-to-face conversation turns out to be highly *interactive* in that it displays “frequent occurrence of features like first and second person pronouns, questions, hedges, contractions, *that*-clauses, *if*-clauses” (Biber and Finegan 1986:40); highly *situated*, given its frequent use of place and time adverbs; and *immediate* in that it has more present than past tenses. Multi-Dimensional

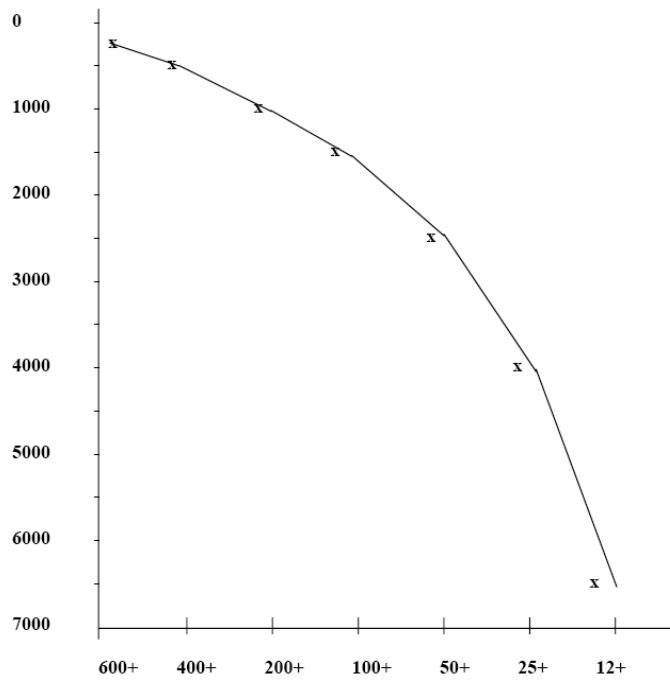
³⁸ Factor analysis “empirically identifies the groups of co-occurring linguistic features and provides the basis for the interpretation of the underlying textual dimensions in a given domain” (Biber and Finegan 1986:23), whereas cluster analysis “empirically identifies the groups of texts that are maximally similar in their exploitation of the textual dimensions, providing the basis for interpretation of these groupings as text types” (Biber and Finegan 1986:23). For further details see Biber (1988) and Chapter 5.

analysis has also demonstrated that the spoken language, in particular face-to-face conversation, is interpersonal and affective (cf. Biber's 1988 Dimension 1); with non-narrative concern (cf. Biber's 1988 Dimension 2); situation-dependent (cf. Biber's 1988 Dimension 3); not particularly persuasive (cf. Biber's 1988 Dimension 4); and with non-abstract information (cf. Biber's 1988 Dimension 5).

Interestingly, McCarthy (1999:2) claims that probably a basic or 'core' vocabulary of spoken English exists. The basis of his claim is the fact that "in computer-based frequency counts, there is usually a point where frequency drops off rather sharply, from hard-working words which are of extremely high frequency to words that occur relatively infrequently". Figure 4 illustrates this phenomenon: the vertical axis of the graph shows how many words in the CANCODE corpus³⁹ actually occur at the given frequencies, whereas the horizontal axis shows frequency of occurrence, for example, 600+ indicates words occurring more than 600 times in the sample, 400+ words occurring more than 400 times, etc. (McCarthy 1999:3). This offers further evidence of speakers opting for repetitive structure when talking in that it is evident that "round about 2000 words down in the frequency ratings, the graph begins to drop more steeply, with a marked decrease in the number of words that occur more than 100 times (this occurs most noticeably from word number 1500 onwards in the list)" (McCarthy 1999:5). It can then be concluded that "words occurring approximately 100 times or more in this sample belong to some sort of heavy-duty core vocabulary, amounting to about 1500 words" and the kind of words that this core vocabulary seems to embrace are articles, pronouns, auxiliary verbs, demonstratives, basic conjunctions, etc., namely, grammar/function words (McCarthy 1999:5).

³⁹ CANCODE stands for "Cambridge and Nottingham Corpus of Discourse in English. The corpus was established at the Department of English Studies, University of Nottingham, UK, and is funded by Cambridge University Press" (McCarthy 1999:2).

Figure 4. Decline in frequency of occurrence of words in the CANCODE sample (McCarthy 1999:3)



McCarthy (1999) identifies nine broad categories of a basic spoken vocabulary, and discourse markers (e.g. *I mean, right, well, so, good, you know, anyway*) are among them. These categories also include modal items, (e.g. modal verbs, lexical modals, adverbs, and adjectives), delexical verbs (such as *do, make, take, and get*) interactive words (e.g. *just, whatever, thing(s), a bit, slightly, actually, basically, really, pretty, quite, literally*), basic nouns (e.g. *person, problem, life, noise, situation, sort, trouble, family, kids, room, car, school, door, water, house, TV, ticket*), general deictics (e.g. *this, that, here, there, now, then, ago, away, front, side, ...*), basic adjectives (e.g. *lovely, nice, different, good, bad, horrible, terrible, different*), basic adverbs (especially those referring to time like *today, yesterday, tomorrow, eventually, finally*; frequency and habituality, like *usually, normally, generally*; and manner and degree like *quickly, suddenly, fast, totally, especially*), and basic verbs for actions and events (e.g. *sit, give, say, leave, stop, help, feel, put, listen, explain, love, eat, enjoy*).

All the features illustrated so far make spontaneous conversation more informal and less influenced “by the traditions of prestige and correctness often associated with publicly available written texts, where the English language is ‘on its best behaviour’” (Biber *et al.* 1999:1050). Besides, spontaneous conversation usually takes place between people who

know one another and, consequently, may employ a vernacular range of expressions which favors informality. Biber *et al.* (1999:1050) give some examples of forms such as *ain't*, *y'all* and *me and Ann* retrieved from conversational data which tend to be regarded as non standard forms.

Two observations are worth underlining though. First, the intrinsic pragmatic nature of spontaneous conversation should not be treated as “performance error” (Miller and Weinert 1998:23); indeed, the features of spontaneous conversation simply reflect the conditions under which it is produced (Miller and Weinert 1998:23): “the structures of spontaneous spoken language have developed in such a way that they *can* be used in the circumstances in which conversation [...] usually takes place.” (Miller and Weinert 1998:23). That is to say that “language is potentially sensitive to all of the contexts in which it occurs [...] and **reflects** [bold in text] these contexts because it helps to constitute them” (Schiffrin 1987:5). Second, as pointed out by Halliday (1992), spoken language appears to lack a clearly defined shape or form only if it is perceived in terms of its transcription: if written texts were transcribed including all their planning processes, they would appear amorphous as well.

2.3 Discourse Markers

As illustrated in Section 2.1.2, discourse markers (henceforth DMs) – a term which Siepmann (2005:37) attributes to Labov and Fanshel (1977:156)⁴⁰ – play an important role, with other similar devices, both in signaling the dynamic and interactive role of the speaker's utterance and in keeping the conversation going. Besides, as pointed out by McCarthy (1999), DMs (e.g. *I mean, right, well, so, good, you know, anyway*) are among the nine broad categories of the basic spoken vocabulary he has identified.

Similarly, Kennedy (1998) and Biber *et al.* (1999) maintain that *you know* is very frequent in conversation, and Erman (2001:1353) points out that *know* is another member of the core vocabulary of the English language and that *you know*, especially, seems to have a very special status in that it is “one of those frequent combinations of words, which, forming part of larger prefabricated structures, in turn form lexicalized sentence-stems and can be expanded and changed in various ways”. Table 4 below illustrates chunks of *you know* in larger

⁴⁰ Labov and Fanshel (1977:156) use the term *discourse marker* to describe *well* pointing out that it “refers back to some topic that is already shared knowledge among participants”.

prefabricated structures from Erman's data⁴¹ (2001:1353):

Table 4. *You know* in larger prefabricated structures from Erman (2001:1353)

***You know* in prefabricated structures**

Structure	
do/if you know what I mean (like)	33
you know + NP (proper name, proform, etc.)	15
you know how (like) + S	9
you know what ((+NP often proper name) done (does, did))	9
you know (the bit) (like) when + S	7
like you know	3
I mean you know	2
you know like	2
it's like you know when + S	2
you know why + S	1
you know whatever	1
guess what? you know + S	1
you know the thing that gets me	1
like cos you know like	1
you know just sort of like	1
you know I mean	1
Total	89

This very special status is also highlighted by the following quote from Crystal (1988), which well illustrates both the importance and use of DMs in general, which he calls *parenthetical phrases*, and of *you know*, in particular, in spontaneous speech production:

You know, and other parenthetical phrases of English, are really far more complex and important than we usually allow. I tend to think of them as the oil which helps us perform the complex task of spontaneous speech production and interaction smoothly and efficiently. They give the speaker and opportunity to check back, to plan ahead, and to obtain listener reaction. They give the listener an opportunity to keep up and to react. If we all had perfect self-control, memory, attention, and logical thought process, doubtless they would be unnecessary; but we haven't. We may admire those who approach this idea state, and who can speak without a trace of non-fluency. But language was never intended to be restricted to an elite corps. (Crystal 1988:48).

⁴¹ Erman's (2001:1338) data come from the *Bergen Corpus of London Teenager Language*.

2.3.1 Terminology, Classification, and Approaches

Although numerous studies deal with DMS, there is little consensus about their terminology and classification. Such devices, indeed, are known by a variety of names: pragmatic expressions (Östman 1981), pragmatic particles (Östman 1981), discourse markers (Schiffrin 1987; Fraser 1999; Blakemore 2002), discourse connectives (Blakemore 1992), phatic connectives (Bazzanella 1990) discourse particles (Aijmer 2002), parenthetic phrases (Crystal 1988), discourse signaling devices, indicating devices, pragmatic connectives, pragmatic formatives, pragmatic operators, semantic conjuncts, sentence connectives (cf. Schourup 1999 for a detailed account), *inter alia*. It seems that the discussion on terminology is, in particular, about the *discourse particle* versus the *discourse marker* label (Fischer 2006b), the label *discourse marker* being the most accepted, despite the fact that both terms arouse controversy.

The label *discourse particle*, indeed, is probably too narrow: first of all, because it suggests “small, uninflected words that are only loosely integrated into the sentence structure” (Fischer 2006b:4); second, because it is misleading, in that “what is expressed by particles in one language may be expressed by very heavy speech formulae in another” (Fischer 2006b:5). The label *discourse marker*, instead, which is usually preferred because it is considered more inclusive and more functional, may be problematic since the functions that a DM may cover can also be covered by other items: linking functions, for instance, can also be expressed by conjunctions and speech formulas; conversational management functions may also be fulfilled by speech formulas and nonlexicalized metalinguistic devices (Fischer 2006b); and stance can be expressed by modal verbs, adverbs, parenthetic clauses, *inter alia* (Fischer 2006b).

This disagreement among scholars is an issue of primary importance, since it does not only imply labeling a set of linguistic features, but rather involves determining the class they belong to (Pons Bordería 2006). Indeed, not only have most of the terms used to describe them been objected to (cf. Aijmer 2002), but their classification is also particularly problematic. As an example of this disagreement, Schiffrin (1987) considers *you know* as a discourse marker, whereas Fraser (1993; 1999) excludes it from the category. This happens especially because, unlike Schiffrin (1987), Fraser (1999, 2006) classifies pragmatic markers into four main classes – basic markers (which signal the illocutionary force of the basic

message), commentary markers (which signal a message which comments on the basic message), parallel markers (which signal a message in addition to the basic message), and discourse markers (which signal the relationship of the basic message to the foregoing discourse) – which, as illustrated in Tables 5-8, are further sub-classified. Specifically, according to Fraser (1993:10-11), *you know* is a *parallel marker* which “signals a message requesting that the hearer appreciate and/or be in sympathy with the speaker’s point of view”, like *come on*.

Table 5. Fraser’s (2006) Basic Pragmatic Marker Classification

BASIC PRAGMATIC MARKERS (FRASER 2006)	
FUNCTION	EXAMPLES
Basic Pragmatic Markers signal the type of message (the illocutionary force) the speaker intends to convey in the utterance of the segment	a) I promise that I will be on time. b) Please , sit down. [a request but not a suggestion or an order] c) My complaint is that you are always rude.

Table 6. Fraser’s (2006) Commentary Pragmatic Marker Classification

COMMENTARY PRAGMATIC MARKERS (FRASER 2006)		
FUNCTION	SUB-TYPE	EXAMPLES
Commentary Pragmatic Markers signal a message separate from but in the nature of a comment on the basis message	Assessment Markers	Mary hurried as fast as she could, but sadly , she arrived too late for the movie.
	Manner-of-speaking Markers	A: Mark, you’ve got to do something. B: Frankly Harry, I don’t know what to do.
	Evidential Markers	A: Will he go? B: Certainly , he will go.
	Hearsay Markers	A: Is the game still on? B: Reportedly , the game was postponed because of rain.

Table 7. Fraser's (2006) Parallel Pragmatic Marker Classification

PARALLEL PRAGMATIC MARKERS (FRASER 2006)		
FUNCTION	SUB-TYPE	EXAMPLES
Parallel Pragmatic Markers signal a message separate from the basis message	Deference Markers	a) Sir , you must listen to me.
	Conversational Management Markers	a) Now , where were we when we were interrupted? b) Well , we could do it either of two ways. c) Ok , what do we do now?

As for the class of discourse markers – which is an elaboration of Fraser (1999), who divides discourse markers into two main categories depending on whether they relate messages or topics – Fraser (2006) describes four types of markers: contrastive (e.g. *but*), elaborative (e.g. *and*), inferential (e.g. *so*), and temporal (e.g. *then*). Fraser's (2006) classification is illustrated in Table 8 below.

Table 8. Fraser's (2006) Discourse Pragmatic Marker Classification

DISCOURSE MARKERS (FRASER 2006)		
FUNCTION	SUB-TYPE	EXAMPLES
Discourse Markers signal a relation between the discourse segment which hosts them, and the prior discourse segment	CONTRASTIVE MARKERS	<i>but, alternatively, although, contrariwise, contrary to expectations, conversely, despite (this/that), even so, however, in spite of (this/that), in comparison (with this/that), in contrast (to this/that), instead (of this/that), nevertheless, nonetheless, (this/that point), notwithstanding, on the other hand, on the contrary, rather (than this/that), regardless (of this/that), still, though, whereas, yet</i>
	ELABORATIVE MARKERS	<i>and, above all, also, alternatively, analogously, besides, by the same token, correspondingly, equally, for example, for instance, further(more), in addition, in other words, in particular, likewise, more accurately, more importantly, more precisely, more to the point, moreover, on that basis, on top of it all, or, otherwise, rather, similarly, that is (to say)</i>
	INFERENTIAL MARKERS	<i>so, after all, all things considered, as a conclusion, as a consequence (of this/that), as a result (of this/that), because (of this/that), consequently, for this/that reason, hence, it follows that, accordingly, in this/that/any case, on this/that condition, on these/those grounds, then, therefore, thus</i>
	TEMPORAL MARKERS	<i>then, after, as soon as, before, eventually, finally, first, immediately afterwards, meantime, meanwhile, originally, second, subsequently, when</i>

Conversely, Schiffrin (1987:31) provides a broader definition of markers as “**sequentially dependent** [bold in the text] elements which bracket units of talk”. Another factor which may imply differences in categorization is that Schiffrin (1987) includes only initial markers in her categorization, and excludes medial and final ones since, they are not dependent on either prior and upcoming discourse.

Another source of disagreement is the fact that DMs are usually studied from different perspectives, especially because terminological issues usually mirror conceptual distinctions (Fischer 2006b). The following quote outlines the problematic areas:

There are very many studies of discourse particles on the market, and by now it is almost impossible to find one's way through this jungle of publications. For a newcomer to the field, it is furthermore often very difficult to find the bits and pieces that constitute an original model of the meanings and functions of discourse particles. Moreover, the studies available so far are hardly comparable; the approaches vary with respect to very many different aspects: the language(s) under consideration, the items taken into account, the terminology used, the functions considered, the problems focussed on, and the methodologies employed (Fischer 2006b:1).

There are some approaches, for example, that define DMs by means of the property of integratedness they display, namely, as items that constitute parts of utterances or sentences: Fraser (2006), for instance, assumes that there are discourse segments that host DMs; similarly, Lewis (2006) mentions syntactic hosts, while Ler Soon Lay (2006) talks about utterances in which they occur, and Hansen (2006) illustrates discourse particles as instructions to the hearer on how to integrate host utterances into a developing model of the discourse. Thus, they all consider DMs as elements which occur in some host utterances (Fischer 2006b). Conversely, there are a number of researchers who consider them as items that constitute utterances themselves, that is, by means of the property of unintegratedness (Fischer 2006a): Schiffarin (2006) and Yang (2006) consider DMs syntactically detachable and syntactically independent, respectively; Travis (2006) and Fischer (2006a) define them as syntactically, semantically and often prosodically unintegrated; and Diewald (2006) mentions grammatical unintegratedness. Proponents of this unintegrated view of DMs usually focus on the roles DMs may play in the management of conversation and "concern domains such as the sequential structure of the dialogue, the turn-taking system, speech management, interpersonal management, the topic structure, and participation frameworks" (Fischer 2006b:9)

The polyfunctionality of DMs is another factor which splits the literature into further different approaches: there is "a considerable spectrum of possible ways of dealing with the problem of bridging the gap between the single phonological/orthographic form and the many different possible interpretations associated with this form" (Fischer 2006b:12). One of these ways, i.e the *monosemy approach*, maintains that there is a single meaning of DMs

that may be instantiated in context (cf. Fraser 2006, Schiffрин 2006, Travis 2006):

Monosemy: Each phonological/orthographic form is associated with a single invariant meaning. This invariant meaning may describe the common core of the occurrences of the item under consideration, its prototype, or an instruction. Individual interpretations arise from general pragmatic processes and are not attributed to the item itself (Fischer 2006:13).

Another one, i.e. the *homonymy approach*, recognizes the different interpretations of DMs, yet it assumes that they are not related:

Homonymy: There are a number of readings that are identifiable as distinct. No relationship between the different readings is assumed, and the different senses are described in numbered or unnumbered lists, sometimes associated with their conditions of usage, such as, for instance, the structural contexts in which they occur (Fischer 2006b:13).

In between these two opposite approaches, there are others, like the *polysemy* and the *polysemy in the narrow sense* approach, which assume that there is not a single invariant meaning component of DMs, but rather different distinct readings which are related (cf. Hansen 2006, Lewis 2006, Travis 2006). More specifically, the *polysemy* approach claims that “a single phonological/orthographic form may be used with a number of different, recognisable interpretations that are assumed to be related” (Fischer 2006b:13), whereas the *polysemy in the narrow sense* approach maintains that “a single phonological/orthographic form is associated with a number of distinct readings that are related by a set of general relationships. These readings do not necessarily share common meaning aspects” (Fischer 2006b:13). Besides, as pointed out by Fischer (2006b), there is also “a broad spectrum of models that take the monosemy approach as a starting point but that furthermore attempt to account for the different senses observable by providing models of mechanisms that relate the invariant meaning to the distinct but motivated readings” (Fischer 2006b:13).

Other differences in terms of categorization may also be ascribed to the slant of the approach followed: Schiffрин (1987, 2001), for instance, offers a discourse-based approach,

which describes DMs as working at different levels of discourse and contributing to discourse coherence by connecting utterances; Fraser (1990, 1993, 1999), who includes in his categorization a very different set of items (e.g. *linking adverbs*, for instance, cf. Table 8), proposes a pragmatic approach which sees DMs as part of the grammar and members of a pragmatic, and not syntactic, category Sperber and Wilson (1986, 1993) and Blakemore (1987), on the other hand, focus on relevance by considering DMs (or rather, discourse connectives) as expressions which contribute to relevance by guiding the hearer towards the intended contextual effects, hence reducing the overall effort required.

2.3.2 General Traits of Discourse Markers

The scarce consensus about the terminology and classification of discourse markers shows the extent to which they are intriguing objects of study. The interest in them may well be due to the fact that “they promise the researcher ready access to the very fabric of talk-in-progress” (Redeker 2006:339) and the continuity of conversation (Erman 1987, Biber *et al.* 1999, Aijmer and Stenström 2005). DMs, indeed, play an active part both in characterizing conversation and in keeping it going (cf. Section 2.2).

Despite the different labels applied, it clearly emerges from the literature that DMs can be identified on the basis of some distinct features that they constantly display: connectivity, optionality, reinforcement / facilitation, contextuality, functional-pragmatic nature and multi-categoriality, multi-functionality, initiality, and orality.

DMs are usually described as devices employed to connect utterances or other discourse units (Schourup 1999, Redeker 2006): Schiffрин defines them as “sequentially dependent elements which bracket units of talk” (1987:31); Erman (1987:77) as “connective elements”; Fraser (1996:186) as “expressions which signal the relationship of the basic message to the foregoing discourse”; Hansen (1997:160) as “linguistic items of variable scope, and whose primary function is connective”; Fraser (1999:938) as elements that “impose a relationship between some aspect of the discourse segment they are part of”; Fuller (2003:25) as markers of “coherence between speakers turns”; and Fraser (2006) as lexical expressions that signal a relationship which exists between adjacent discourse segments. Siepmann (2005:41) points out that DMs “mark a relationship between text spans” and, consequently, “serve to indicate how one unit of discourse is to be constructed in the light of another”

(Siepmann 2005:43).

Another feature which usually emerges from the literature is that DMs are usually considered to be syntactically optional, i.e. “syntactically detachable from the sentence” (Schiffrin 1987:328), in that their removal does not alter the grammaticality of their host sentence (e.g. Fraser 1988, Schourup 1999), or, in Brinton’s (1996:267) words, their omission “renders the text neither ungrammatical nor unintelligible”. However, their optionality does not render them “meaningless decorations” (Aijmer 2002:2) or redundant (Brinton 1996, Schourup 1999). On the contrary, they usually reinforce and guide participant understanding (Redeker 2006) like “processing instructions intended to aid the hearer in integrating the unit hosting the marker into a coherent mental representation of the unfolding discourse” (Hansen 1998:236), or, as Siepmann (2005:44) puts it, they highlight the interpersonal function by “expressing speaker or writer stance or in securing cooperation and understanding”. This facilitation of the listener’s processing task, or focus on the information framed by the specific discourse marker in question (Fuller 2003:27), is another characteristic of DMs about which there is consensus.

In terms of semantics, the only aspect the literature agrees on is that the meaning (and function) of DMs strictly depends on the context they occur in (Hansen 1998, Fraser 1999, Fuller 2003). The most flexible semantic definition is probably the one suggested by Hansen (1998:245) who describes them as linguistic items which “have only a meaning *potential* which must be actualized by a specific hearer in a specific context, via the construction of a mental representation”. In other words, Hansen (1998:245) posits that since language provides a finite means of expressing an infinite number of messages, “most, and perhaps all, linguistic units are therefore inherently variable to some degree, and the actualized meaning of a given item will be influenced by that of the other items with which it co-occurs, and by the grammatical and sequential structure imposed on them”.

A different view is offered by those scholars who, like Schiffrin (1987:127), maintain that some DMs, such as *well* and *oh*, lack an inherent semantic meaning in that they only mark response and are consequently available for a general discourse function. Still others, like Carlson (1984) and Bolinger (1989) totally reject the lack of meaning of DMs maintaining that *well*, in particular, has a *core meaning* related to *acceptance*. In much the same way, Fraser (1990, 2006), Redeker (1991:1165), Schiffrin (1987) herself – who, in fact, excludes *well* and *oh*, as just pointed out – and Hansen (2006), *inter alia*, also maintain that DMs have an

invariant semantic content, usually defined as *core meaning*, or *basic meaning* (cf. Fox Tree and Schrock 2002:736) which restricts the possible interpretations of utterances in which a DM appears (Hansen 2006). Consequently, even “when a particular DM is claimed to be semantically empty, it is usually nevertheless held to have an invariant core of some kind” Schourup (1999:249).

Schourup (1999:242) claims that the lack of meaning posited by Schiffriñ (1987:127) “may be taken to imply only that such DMs contribute nothing to the truth-conditions of the proposition expressed by an utterance” (cf. also Blakemore 2002:12). Similarly, Fraser (1993:4) underlines that DMs do not participate as part of the propositional content of the utterance; indeed, they are detachable and can be deleted without changing the content meaning or the grammaticality of the sentence (cf. also Bazzanella 1990:632). However, their absence “does remove a powerful clue about what commitment the speaker makes regarding the relationship between the basic message conveyed by the present utterance and the prior discourse” (Fraser 1993:4). In particular, according to Fraser (1993:6), the *core meaning* both signals the type of relationship (like change of topic, consequence, parallelism, contrast) “between the current basic message and the prior context” and “provides the starting point for the interpretation of the commentary message”. Conversely, Aijmer (1996:23) maintains that the *core meaning* is “a fairly abstract notion” and she defines DMs in terms of grammaticalization, i.e. as those particles that have been grammaticalized resulting “in a class of words with unique formal, functional and pragmatic properties” (Aijmer 1996:16). A slightly different view is provided by Hansen (1998:238), who maintains that DMs are “items that are still in the *process* [italic in the text] of being grammaticalized” and “the farther they have moved along the grammaticalization cline, the greater variety of function they seem capable of assuming”. The *core meaning* approach also contrasts with the *polysemy approach* (Hansen 2006, Lewis 2006) described above (cf. also Section 2.3.1), which assumes that DMs can have more than one meaning, but the different meanings they can display are related.

On the one hand, there is little agreement as to the class some discourse markers belong to (Schiffriñ 1987:40; Aijmer 1996:7), especially because they do not constitute a word class in the traditional sense due to the fact that there are neither semantic nor morpho-syntactic criteria which delimit it (cf. Siepmann 2005:44; Hansen 1998:236). On the other, it is fairly clear that the category of discourse markers can be accounted for in functional-

pragmatic terms (Hansen 1998; Pons Bordería 2006): DMs are usually defined functionally and said to be independent of syntactic categorization. This implies a rejection of a grammatical paradigm and consequent exclusion of formalist approaches, and an awareness of distinctive functional features as the basis for the description of DMs (Pons Bordería 2006). Indeed, even though syntactically speaking DMs can be categorized as coordinate conjunctions, subordinate conjunctions, prepositions, prepositional phrases, or adverbs (cf. Fraser 2006), they usually display an extrinsic function (Schourup 1999): “categories to which extrinsic DM function has been attributed include adverbs (e.g. *now*, *actually*, *anyway*), coordinating and subordinating conjunctions (e.g. *and*, *but*, *because*), interjections (e.g. *oh*, *gosh*, *boy*), verbs (e.g. *say*, *look*, *see*), and clauses (e.g. *you see*, *I mean*, *you know*), though many would wish to shorten or lengthen this list of categories. When DM status is seen, instead, as a matter of syntactic categorization, multi-categoriality is viewed diachronically and DMs are taken to arise from other categories through historical processes” (Schourup 1999:234).

Another pragmatic trait the literature usually agrees on is the multi-functionality of DMs: “discourse markers have been shown to operate on the syntactic-semantic and pragmatic levels simultaneously” (Siepmann 2005:44; cf. Hansen 1998:238) and the single DMs have been shown to cover a range of distinctive functions in different situations and contexts⁴² “depending on the roles and relationships of the interlocutors” (Fuller 2003:25).

⁴² *I mean*, for instance, is maintained to be used to evaluate (Fox Tree and Schrock 2002, Brinton 2003), to inform (Erman 1987, Fox Tree and Schrock 2002), to be more precise and explicit (Erman 1987, Brinton 2003), to make adjustments (Erman 1987, Fox Tree and Schrock 2002, Brinton 2003), to correct and reformulate previous utterances (Erman 1987, Brinton 2003), to justify and modify (Erman 1987, Fox Tree and Schrock 2002), to make the speaker less committed (Erman 1987, Brinton 2003, Fox Tree and Schrock 2002), to introduce a mitigation (Erman 1987, Tottie 2002), to emphasize (Brinton 2003), and again it is also described as a “softener” (Crystal and Davy 1975), as a turn-taking device or as a turn-yielder (Erman 1987), as a “compromiser” (James 1983), as a hesitation marker (Erman 1987), and as a politeness marker (Fox Tree and Schrock 2002, Brinton 2003).

Similarly, *you know*, is often said to provide a form of rhythmic pattern (Jefferson 1973, Macaulay 2000), to convey additional information or to organize it (Erman 1987), to introduce backgrounded knowledge or parenthetic comments (Erman 1987, Macaulay 2000), to narrow the scope (Erman 1987), to round off the theme (Erman 1987), to introduce an exemplification/clarification (Erman 1987), to mark the end of a syntactic unit/argument (Erman 1987, Macaulay 2000, FoxTree and Schrock 2002), to emphasize (Macaulay 2000, Fox Tree and Schrock 2002) and to draw attention (Macaulay 2000), to relieve the speaker from being completely committed to the truth value of the proposition in question (i.e. with face-saving function, cf. Schourup 1985, Erman 2001, Fox Tree and Schrock 2002), to establish rapport with the listener (Tottie 2002), and to create coherence (Erman 1987, 2001). Besides, *you know* is also described as a “verbal filler” (Brown 1977), as a “fumble” (Edmondson 1981), as “clause internal ‘restarts’” (Schourup 1985), as a “staller” (Erman 1987), as a turn-taking device and as a turn-yielder (Erman 1987), as a confirmation seeker (Erman 1987), as a topic shifter (Erman 1987), as a repair marker (Erman 1987), as a hesitation marker (Erman 1987), as a booster (Holmes

Nevertheless, DMs “can have specialized functions in one particular type of discourse” (Norrick 2001:11). This multi-functionality, or polyfunctionality, according to Pons Bordería (2006), should be interpreted at two levels: first, at a type⁴³ level, in that a DM can convey different values (*but*, for instance, is polyfunctional because it expresses contrast, especially in monological uses, and disagreement, especially in dialogical uses; Pons Bordería 2006); second, at a token level: a token of a DM is polyfunctional if it displays different functions in different discourse levels (a token of *but*, for example, can express contrast in a sentence level and disagreement in an interactional level).

Initiality/initial position is not considered the first criterion for DM status: Erman (1987) and Forchini (*forthcoming*)⁴⁴, for instance, point out that *you know* and *I mean* occur especially in mid-position (cf. also Erman 2001 on *you know* occurring in mid-position especially and Table 10 below) and Fuller’s (2003) data⁴⁵, which is illustrated in Table 9, clearly show how DM position may vary. In much the same way, Bazzanella (1990:634) underlines that DMs can occur in initial, medial, or final position, depending also on individual usage, and acquire different functions according to their occurrence.

1997), as a hedge (Holmes 1997, Erman 2001), as a shared knowledge marker (Erman 2000) and as an approximator (Erman 2001).

⁴³ The term *type* usually refers to the keyword or the target item in a given corpus, whereas the term *token* refers to all the occurrences of a particular *type* in a corpus (cf. Kennedy 1998: 247-251).

⁴⁴ Forchini’s (*forthcoming*) corpus is based on the following movies: *Mission: Impossible II* (J. Woo, 2000); *Shallow Hal* (B. and P. Farrelly, 2001), *Ocean’s Eleven* (S. Soderbergh, 2001); *The Matrix Reloaded* directed by (A. and L. Wachowsky , 2003)

⁴⁵ Fuller’s (2003:30) data “were collected in two speech contexts: research participants were first interviewed by a graduate research assistant, and were then asked to record a conversation with a close friend or family member”.

Table 9. Percentages of DMs used in different syntactic positions from Fuller (2003:41)

DM	Turn initial or 2nd ^a	Turn medial	Utterance medial	Other ^b	Total
<i>Oh</i> —conversation	50 72%	10 15%	0	9 13%	69
<i>Oh</i> —interview	29 58%	17 34%	0	4 8%	50
<i>Well</i> —conversation	46 69%	19 28%	0	2 3%	67
<i>Well</i> —interview	40 55%	31 43%	1 1%	1 1%	73
<i>You know</i> —conversation	7 12%	37 66%	6 11%	6 11%	56
<i>You know</i> —interview	1 1%	132 76%	25 14%	15 9%	173
<i>Like</i> —conversation	1 1%	50 60%	31 38%	1 1%	83
<i>Like</i> —interview	12 6%	95 48%	91 46%	0	198
<i>I mean</i> —conversation	8 23%	24 69%	0	3 8%	35
<i>I mean</i> —interview	10 10%	85 88%	0	2 2%	97
<i>Yeah</i> —conversation	119 72%	14 8%	0	33 20%	166
<i>Yeah</i> —interview	275 61%	56 12%	0	120 27%	451

^a Turn 2nd placement indicates the DM is used following another DM (e.g., *oh, well*) or after a verbal filler (e.g., *uh, um*).

^b Alone or turn final.

However, it seems that DMs may occur predominantly in initial position (Schourup 1999) especially if the term *initial* is broadened slightly and two aspects are taken into account: first, if the tendency toward initiality is not understood to refer to the position of the first word in an utterance, but rather to the position of DMs in relation to the central clause elements (Schourup 1999); second, if their usual tendency to introduce the discourse segments they mark is understood as a function similar to the one they serve initially (Hansen, 1997:156; cf. Schiffarin, 1987:31-32,328). As pointed out by Schourup (1999:233), the tendency to occur in initial position “is probably related to their ‘superordinate’ use to restrict the contextual interpretation of an utterance: in general it will make communicative sense to restrict contexts early before interpretation can run astray”. Besides, another common feature of DMs that confirms (and favors) *initiality* is that they may occur without the presence of the initial part of the sentence/utterance, when the non-linguistic context provides a suitable message (Schourup 1999). The dropping of the initial part of the utterance, indeed, allows the DM to be the very first item in the utterance.

Although a large number of forms described as DMs occur primarily in speech (e.g. *by the way, well, after all*; cf. Schourup 1999), “no principled grounds exist on which to deny DM status to similar items that are largely found in written discourse (e.g. *moreover, consequently, contrariwise*)” (Schourup 1999:234). There are, in fact, numerous points of contact between discourse markers used in speech and writing (Siepmann 2005): for example, “the somewhat cumbersome topic shifter *it may be noted in passing that*, peculiar to written, or written-to-be-

spoken register, may be said to parallel closely the use of *incidentally* in everyday speech or writing” (Siepmann 2005:38). Association of a DM within the written or spoken register is, indeed, “often tied only to the relative formality/informality of the DM (e.g. *also* versus *moreover*). The meaning of a marker may also ally it to one channel or the other. For example, some putative DMs such as *conversely* and *in contrast* encode a high degree of utterance planning” (Schourup 1999:234). This further validates what has already been pointed out in Section 2.2, namely, that the traits of spontaneous conversation simply reflect the conditions under which it is produced (Miller and Weinert 1998:23), and this, of course, can be said for written language as well.

2.3.3 Focus on *You Know*: a Functional Categorization

Although numerous studies deal with expressions like *you know*, as highlighted in Section 2.3, there is little consensus about its terminology, classification and function. As stated above, an example of this disagreement can be seen in Schiffrin (1987) and Fraser (1999): the former includes, whereas the latter excludes *you know* from DMs. Function seems problematic too; however, the disagreement that emerges from the literature is only superficial: indeed, as I demonstrate in the following paragraphs, a closer look and a wider perspective not only proves the multi-functionality of the DM category (cf. Section 2.3), but also allows one to interpret *you know*, in its different shades of meaning, as a feasible case of multi-functionality.

One of the aims of the present study is to provide a clear description of *you know*, of which detailed examples are given within the analyses in Chapter 4. This is done by classing together the interpretations suggested by the literature on *you know* and reorganizing those that make similar statements. Thus, the DM in question may be described as displaying the following functions:

1. The *telling/commenting function*, which signals that the speaker is telling or commenting on something, is evident when *you know* is described by the literature as “an information state marker” (Schiffrin 1987:294) employed “in particular in narrative parts of conversations” (Östman 1981:16) to add (new) information (Erman 1987), to introduce background information or parenthetic comments (Erman 1987; Macaulay 2000), and to shift the topic (Erman 1987).

2. The *turn-dealing function*, which signals that the speaker is taking or passing

the turn, is present in the literature when *you know* is claimed to be used as a turn-taking device (Östman 1981), as a turn-switching marker (Östman 1981), as a turn-yielder (Erman 1987), as a marker of the end of a syntactic unit/argument (Erman 1987, Macaulay 2000, FoxTree and Schrock 2002), or as a confirmation-seeker (Erman 1987) which checks that the listener is understanding what is being said (Crystal 1988).

3. The *emphasizing function*, which signals that the speaker is giving special prominence or drawing attention to something, is envisaged in the literature when *you know* is labeled as a booster (Holmes 1997), as an emphaser (Macaulay 2000, Fox Tree and Schrock 2002), and an attention drawer (Macaulay 2000).

4. The *clarifying function*, which signals that the speaker is making a statement or a situation more comprehensible, is illustrated by the literature when *you know* is considered a strategy to introduce an exemplification or a clarification (Erman 1987), to narrow the scope (Erman 1987), to round off the theme (Erman 1987), to repair (Erman 1987), or to mark the speaker's upcoming modification of the meaning of his/her prior talk (Schiffrin 1987).

5. The *shared knowledge marking function*, which signals that the speaker is appealing for or awakening the knowledge (s)he shares with the listener, is illustrated by Erman (2001), who describes *you know* as a *shared knowledge marker* (cf. also Bazzanella 1990:642).

6. Finally, the *time-stalling function*, which signals that the speaker is trying to find the most appropriate expression either because (s)he does not know what to say or how to say it, is illustrated by the literature when *you know* is described either as a verbal filler (Brown 1977), as a pause-filler (Östman 1981), as a fumble (Edmondson 1981), as a clause internal restart (Schourup 1985), as a staller (Erman 1987), or as a mitigator (Östman 1981), as a hedge (Holmes 1997, Erman 2001), as an approximator (Erman 2001), and as a hesitation marker (Erman 1987) useful to the speaker to be relieved from being completely committed to the truth value of the proposition in question (Schourup 1985, Erman 2001, Fox Tree and Schrock 2002).

Regarding functions in context, Erman (1987; 2001) provides a detailed description of the functions that *you know* acquires according to its turn position. Erman (2001) points out that *you know* occurs with the highest frequency in mid position: as illustrated in Table 10, 77% of occurrences in the Bergen Corpus of London Teenager Language (i.e. COLT) and 84.6% occurrences in the London-Lund Corpus (i.e. LLC) are in mid-position.

Table 10. Erman's (1987) *you know* frequency according to turn position

Table 1
Position in the turn in LLC and COLT

Corpus	Initial		Medial		Final		Total	
	N	%	N	%	N	%	N	%
LLC	14	5.0	236	84.6	29	10.4	279	100
COLT	28	9.9	217	77.0	37	13.1	282	100

In particular, as shown in Table 11, mid position *you know* functions as a topic shifter or as a turn taking device, which, in terms of my functional categorization listed above, means that the *telling* and *turn-dealing functions* are the most frequent functions of *you know* in spontaneous spoken conversation.

Table 11. Erman's (1987) *you know* functions according to turn position

YOU KNOW FUNCTION & POSITION, ERMAN 1987	
INITIAL	REPAIR MAKER STALLER
MEDIAL	TOPIC SHIFTER TURN-TAKING DEVICE
FINAL	CONFIRMATION-SEEKER TURN-YIELDER

Similarly to Erman (1987), though without providing quantitative data, Crystal (1988:47) maintains that when *you know* is used at the beginning of a sentence, it is usually employed to soften, as “a verbal equivalent to a gentle hand on the shoulder”; in terms of the above categorization, it is used with a *time-stalling function*. When *you know* is used at the end of the sentence, Crystal (1988:47) interprets it “as a kind of tag question – as a check that the listener is understanding what is being said” like Erman's (1987) confirmation-seeker; in the above-mentioned terms, it is used with a *turn-dealing function*. Finally, unlike Erman (1987), Crystal (1988:47) posits that when used in the middle, *you know* usually clarifies or amplifies the meaning of the sentence and underlines that “the next words are particularly important”; according to the present categorization, this means that it is used with a clarifying or

emphasizing function.

Interestingly, the highest frequency of the *telling function* allows one to speculate that *you know* is closer to its literal (cf. Schiffarin 1987), rather than non-discourse-marker-like meaning, recalling the semantics of the full verb *to know*, even though it is employed as a pragmatic device that signals a new topic being introduced by the speaker.

Another interesting study, which further highlights the relevance of the *core meaning* of *you know*, is provided by Schiffarin (1987), who unlike Erman (1987)⁴⁶, takes into account only turn-initial *you knows*. Schiffarin (1987:267) points out that in her data “*y'know* marks transitions in information state which are relevant for participation framework”; indeed “*y'know* functions within the information state of talk” (Schiffarin 1987:267). More specifically, Schiffarin (1987) claims that the literal meaning of *you know* (which “refers to the cognitive state in which one has information about something”; Schiffarin 1987:267) suggests its uses in information states via two functions: as “a marker of meta-knowledge about what speaker and hearer share” and as “a marker of meta-knowledge about what is generally known” (Schiffarin 1987:268) working “basically within the information state of the talk” (Schiffarin 1987:309).

2.4 Summary of the Present Framework

In the present work, these particles/markers, including the one which will be investigated in detail, i.e. *you know*, are called *discourse markers* on the ground that they belong to discourse (either spoken or written, cf. Siepmann 2005:37) and, presumably, mark something by “focus[ing] on the organization and orientation of the discourse” (Erman 1987:128). Similarly, Travis (2006) uses the term to refer to those items that act on, or mark, segments of discourse. This rather simplistic justification not only avoids adding to “the flood of terminology threatening to submerge language science” (Siepmann 2005:37), but it also aims to be inclusive, rather than exclusive, and matches a theoretical framework which starts from frequency (i.e. presence vs. absence in the data) and ends up by considering the role of the item in the environment within which it occurs (i.e. it is function in context). The present

⁴⁶ It is worth noting that in Erman (1987), initial *you know* is found as a *repair maker* and a *staller* (cf. Figure 9), which in terms of my re-categorization means it has a *time-stalling function* and not a *telling* one like in Schiffarin (1987).

approach, close to Schiffriin's (1987), can thus be described as functional, and DMs are defined in the present study as *meaningful pragmatic devices, syntactically detachable from the utterance, which connect discourse units by looking forward and/or backward in discourse and which guide the interpretation of the utterance by marking and reinforcing it.*

As stated above, one of the aims of the present study is to provide a description of *you know*, which can help clarify the disagreement among scholars. Consequently, the different interpretations suggested by the literature on *you know* will be classed together and the DM will be described, if/when the evidence from the data allows it, according to the telling/commenting, turn-dealing, emphasizing, clarifying, shared knowledge marking, or time-stalling function. More specifically, the telling function will be applied when the speaker uses *you know* in contexts where (s)he says something, or comments on it, providing new information or information that (s)he thinks may be unknown to the listener, and sometimes making sure that the (s)he is following the topic of the utterance; the *clarifying function* will be applied when the speaker uses *you know* in contexts where (s)he makes a statement or a situation more comprehensible either by narrowing and specifying what (s)he means or by enlarging and providing further explanation about the topic; the *knowledge marking function* will be applied when the speaker uses *you know* in contexts where (s)he appeals for or awakes the knowledge (s)he shares with the listener; and the *time-stalling function* will be applied when the speaker uses *you know* in contexts where (s)he tries to find the most appropriate expression either because (s)he does not know what to say or how to say it, both under unpleasant and pleasant circumstances where the situation is embarrassing (e.g. like when somebody pays a compliment or invites somebody out and is afraid of a negative response). So as to avoid subjective interpretations, *you know* will be labeled as occurring with a *time-stalling function*, only when it occurs with a negative semantic prosody or with other discourse markers, inserts, repetitions, syntactic blends which clearly imply some hedging or hesitation.

CHAPTER 3. MOVIE CONVERSATION

After the description of the typical traits of face-to-face conversation in Chapter 2, the present chapter examines the second domain in which Multi-Dimensional studies will be applied and *you know* will be analyzed (cf. Chapter 4), namely American movie conversation. This type of domain is usually considered by the literature as a kind of prefabricated speech which is written to sound like authentic speech (Sinclair 2004b, Taylor 1999, Rossi 2003, Pavesi 2005). Indeed, it is spoken by actors who have to follow a planned and written script, yet it has to appear spontaneous within its artificial settings. The aim of the present chapter is thus to focus on its linguistic features so as to compare and contrast them to those of spontaneous spoken conversation (cf. Chapter 2).

The starting point for focusing on movie language is the fact that despite the large number of studies on dubbing and subtitling (cf. Baccolini and Bollettieri Bosinelli 1994; Pavesi 1994, 2005; Bollettieri Bosinelli 1998; Pavesi and Malinverno 2000; Taylor 2000a, 2000b, 2000c, 2003; Gottlieb and Gambier 2001; Taylor and Baldry 2004; Bruti and Perego 2005; Bruti 2006), the actual language of movies has not received much attention (cf. Taylor 1999:247; Rossi 2003:93). Furthermore, the vast majority of existing research concentrates on movie web scripts, rather than actual movie dialogs, (cf. Taylor 1999:262).

This apparent lack of interest and description can be ascribed to a number of possible reasons. First of all, a large amount of channels and codes are involved; it is not always easy to de-codify and interpret all of them, for they co-exist and inter-play simultaneously, transmitting and influencing the meaning and pragmatics of the message delivered. As a consequence, when analyses happen to focus on only one of these channels and codes (on the auditory channel or on the linguistic code, for instance), they necessarily imply some pragmatic and semantic loss. But these channels and codes, being pragmatically and semantically interlinked, should not be treated separately.

The second reason depends on the relative difficulty of finding transcriptions of movie dialogs (Rossi 2003:93): as described in Chapter 1 on methodology, the absence of movie corpora does not help in this regard. Furthermore, the scripts which are easily downloadable from the web differ considerably from what is actually said in movies; consequently, they cannot be considered as representative of movie conversation. One way to solve this problem is to manually transcribe movie dialogs, even though manual transcription

is an extremely time-consuming process; and this is, undoubtedly, another limiting factor. Another reason that might not have favored movie language investigation is the prejudice against this kind of conversational domain (cf. Rossi 2003:93). It is worth emphasizing, though, that claims that might have discouraged movie language study have been put forward exclusively to point out that movie language does not provide evidence for spontaneous conversation due to its planned and, consequently, non-spontaneous nature; however, this does not imply that movie language cannot be investigated as a variety of its own or that it counts less. The following quote from Sinclair (2004b), already mentioned in the introduction, clearly illustrates this point (my highlighting):

If it is impossible in an early stage of a project to collect the spoken language, then there is a temptation to collect film scripts, drama texts, etc., as if they would in some way make up for this deficiency. They have a very limited value in a general corpus, because they are ‘considered’ language, **written to simulate speech in artificial settings. Each has its own distinctive features**, but **none truly reflects natural conversation**, which for many people is the quintessence of the spoken language. [...] **such records are not likely to be representative of the general usage of conversation** (Sinclair 2004b:80).

Sinclair’s (2004b:80) quote is fundamental for the present research for two reasons: first, because it strengthens the value of searching for the distinctive features of movie language; second, because it openly declares that the two conversational domains differ and that movie language has “a very limited value” because it does not reflect natural conversation and, consequently, is “not likely to be representative of the general usage of conversation”. The crucial missing element in the comment is, however, empirical evidence.

With this premise, movie investigations become, then, extremely interesting for the description of “media language in its own right” (Mansfield 2006:17), which, of course, is also influenced by its multiple channels. As regards the present investigation, Section 3.1 illustrates the multi-modal aspect of movies, whereas Section 3.2 focuses on the linguistic code, highlighting the co-presence of fictitious and spontaneous traits in movie language. More specifically, Section 3.2.1 points out the non-spontaneous elements that characterize movie conversation and distinguish it from spontaneous spoken conversation and Section

3.2.1.1 offers an overview of the reasons for their necessary existence. Section 3.2.2, instead, focuses on the role of language and of the scriptwriter, whose role and decisions are fundamental in the attempt to make movie language sound spontaneous. In particular, Section 3.2.2.1 points out the common linguistic features, usually mentioned by the literature, which give movie language a spontaneous slant, while 3.2.2.2 gives details of one of the most frequent discourse markers in spontaneous conversation, namely, *you know*, so as to see whether it is also present in movie conversation and, if so, to what extent.

3.1 Multiplicity of Channels, Codes, and Messages

Movies are an extremely interesting case of multimedia communication both for the multiplicity of the channels and codes involved and for their mutual interaction, which creates values and meanings that go beyond the mere sum of the parts (Chaume 2004c). Movies can be defined as multimedia products: “multimedia is referred to when speaking of the processing and presentation of text, graphics and pictures, if not animation and motion video” (Cattrysse 2001:1) and movies make meaning “through the use of words, gestures, sounds, music and pictures”, thus acquiring “an audio-visual textuality” (Taylor 1999:265). Apart from this, interactivity of the parts is another important parameter that is generally required of multimedia products (Cattrysse 2001:1) and one of the most important factors of movies is that their channels and codes are not independent; rather, they interact with one another simultaneously.

As Chaume (2004a:14) and Pavesi (2005:9) point out, movies offer two channels of communication: the auditory (i.e. the acoustic side of the movie, e.g. the language/languages, sounds, noises, etc.) and the visual channel (i.e. the image side of the movie, e.g. road signs, shop fronts, clothes, colors, body movements, face expressions, etc.). The possible interplays and combinations of these two channels are fundamental to the delivery of their “aural-verbal, aural non-verbal, visual-verbal and visual non-verbal messages” (Remael 2001:14).

Furthermore, the production of the meaning of these messages also depends on several signifying codes that work at the same time. According to (Chaume 2004c:16-21), the following codes can be distinguished :

1. *the linguistic code*, is the language used. Its peculiarity lies in the fact that this type of

language has to appear spoken and spontaneous, while, in fact, it is written and planned for artificial settings (cf. Gregory and Carroll 1979, Sinclair 2004b, Pavese 2005 and Section 3.2 which gives a more detailed account of this);

2. the *paralinguistic code*, which denotes features which provide non-verbal information, but which are still auditory (e.g. laughter);

3/4. the *musical* and the *special effects code*, which are represented by songs and illusions created by props, camerawork, computer graphics, etc., that appear in movies;

5. the *sound arrangement code*, which deals with features which either belong to the story (i.e. diegetic sound) or to a person or object which is not part of the story, such as an off-screen narrator (i.e. non-diegetic sound). The *sound arrangement code* implies both the sounds that are produced on-screen (i.e. those associated with the vision of the sound source) or off-screen (i.e. those whose origin is not present in the frame and therefore not visible simultaneously with the perception of the sound);

6. the *iconographic code*, which represents the icons, indices, and symbols in the movie;

7. the *photographic code*, which deals with changes in lighting, in perspective, or in the use of color (e.g. color vs. black and white or intentional use of some colors);

8. the *planning code*, which depends on the types of shots (i.e. close-ups and extreme);

9. the *mobility code*, which includes proxemic (i.e. related to space) and kinetic (i.e. related to motion) signs, and the screen characters' mouth articulation;

10. the *graphic code*, which is the written language present on screen (i.e. titles, intertitles, texts, and subtitles);

11. the *syntactic code*, which is the editing, namely, the process of shot associations.

The linguistic, the paralinguistic, the musical, the special effects, and the sound arrangement codes (i.e. codes 1-5) are transmitted by the auditory/acoustic channel, whereas the iconographic, photographic, the planning, the mobility, the graphic, and the syntactic/editing codes (i.e. codes 6-11) are transmitted by the visual channel (Chaume 2004c:16). As stated above, all these codes and channels are not independent, but interact simultaneously. This interaction is what makes movie language difficult to analyze; indeed, it is not always possible to take into account the whole interplay of these codes and channels and the most natural consequence which may emerge is the loss of some pragmatic and semantic features.

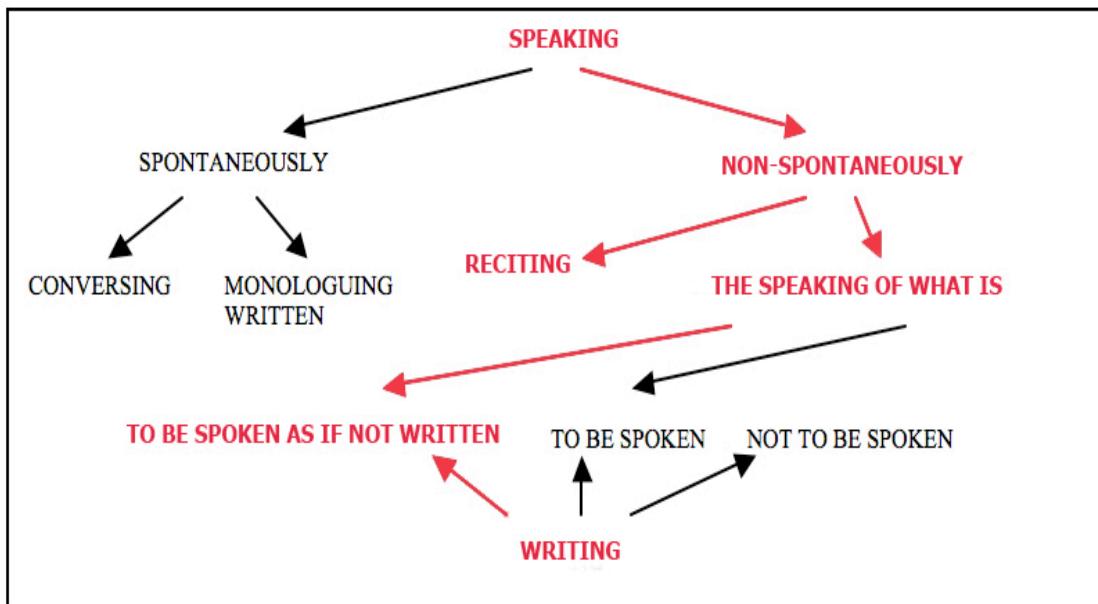
3.2 Fictitiousness and Spontaneity

The status of movie language is rather complicated and paradoxical. In terms of spontaneity, movie language is a fake: it is planned and artificial, yet it pretends to be authentic (Sinclair 2004b, Taylor 1999, Rossi 2003, Pavesi 2005), and displays some traits typical of spontaneous spoken conversation. Nevertheless, there are factors that tie it to its non-spontaneity (cf. Nencioni 1976 and next section). From the opposite perspective, however, it can also be claimed that in spite of being planned and artificial, movie language does resemble spontaneous spoken conversation because it displays some spontaneous traits (Nencioni 1976, Taylor 1999, Rossi 2003, Pavesi 2005). Interestingly, whatever perspective may be taken, it is undeniable that movie language is a variety of its own (presumably with different sub-genres) and, in this respect, it cannot be claimed to be a fake. Besides, it is evident that fictitiousness and spontaneity are features that can (and do) co-exist in movies. Of course, movie speech can also be labeled “quasi-speech” as Sinclair (2004b:80) calls it, if the term “speech” is identified with spontaneous spoken language; indeed, movie language is a spoken variety which is written and planned to sound (i.e. to be spoken as if it were) authentic. Consequently, it cannot be said to be 100% spoken and spontaneity is just an illusion. However, it is worth emphasizing that “the fact that actors and directors are able to create this illusion is itself an indication that spontaneous conversation is a recognizable phenomenon with identifiable features” (Taylor 1999:247). It becomes, then, interesting to investigate these features, to find out to what extent and why movie language sounds either authentic or false, and to see how it diverges from other conversational domains. The following sections try to answer these questions by focusing, first, on the non-spontaneous and, then, on the spontaneous nature of movie language.

3.2.1 Non-Spontaneous Spoken Conversation

In terms of the categorization based on Gregory and Carroll (1978) which was illustrated in Section 2.1, there are two main types of non-spontaneous conversation: (1) *reciting* and (2) *speaking of what is to be spoken as if not written*, of what is *to be spoken*, or of what is *not to be spoken*. These are all non-spontaneous domains in that they imply some planning and are all based on some *writing*. The two types are illustrated in Figure 5 below.

Figure 5. Non-spontaneous Conversation (adapted from Gregory and Carroll 1978)



Non-spontaneous conversation is an extremely broad genre which, of course, includes further sub-genres such as “political speeches, lecture notes, and [...] film scripts”, *inter alia* (Taylor 1999:262). In terms of movie language, the focus of this chapter, it is a diamesic variety that can be categorized as non-spontaneous especially for three reasons: first, because it is planned and prefabricated; second, because it is written, or rather, it is written-to-be-spoken as if it were not written, as usually pointed out by the literature (cf. Nencioni 1976; Gregory and Carroll 1978; Taylor 1999; Pavesi 2005); third, because it always implies some reciting, i.e. the speakers are actors who recite and have to follow a script, or screenplay, even when they are asked to improvise (Nencioni 1976, cf. Rossi 2003:94 on actors in Rosselini’s movies). Although all these features imply that movie language should not be studied to describe spontaneous conversation (cf. Sinclair 2004b:80), and that it should be studied to describe features of those languages which are written-to-be-spoken only to see to what extent they diverge from those which are spontaneous, Chapter 4 will give empirical evidence about the fact that it may, instead, have “the potential to provide researchers and teachers with a convenient source of spoken language data” (Quaglio and Biber 2006: 717).

3.2.1.1 Movie Constraints

Apart from the fact that movie language is non-spontaneous, being prefabricated, planned, written-to-be-spoken, and it implies some reciting, there are also other factors, like the time and space constraints of movies (Taylor 1999:265) and the need to sell the movie which tie movie language to its non-spontaneity.

The first factor, the movie length, is a constraint that leads to fictitiousness, in that it obliges the scenes and language to be explicit and compact: there is not much time and space for redundancy in a two-hour movie, consequently, exchanges and scenes must be relevant and precise. Even scenes that might be of interest are often edited out because of the lack of space and time, although they are sometimes inserted in extra sections of the final versions of DVDs.

The second factor, the necessity to sell the movie, needs to be taken into account because it influences (and justifies) the linguistic choices made when building up the dialog. Indeed, this factor is strictly linked to two other important constraints, “the need to relate interesting, exciting or engaging stories” (Taylor 1999:265), and the need to prevent the audience from losing track of the plot, which sacrifices the spontaneity of language. These constraints make the spontaneity of language secondary because, in the interest of box office sales, the story line needs to be involving and clear. strategy to achieve involvement and clearness is seen in the “excess of highly pertinent, dramatic or intriguing exchanges” (Taylor 1999:265) of dialogs which are never “extremely garbled” (Taylor 1999:266). As an inevitable consequence, movie dialogs lose some spontaneity and acquire artificiality: even when the audience is introduced to a scene that starts mid-conversation, for example, which is supposed to recall spontaneous speech, the information exchange is always “artificially clear” (Taylor 1999:267; cf. also Pavesi 2005:34). The same happens when an on-going conversation is stopped by the introduction of a new scene and then re-presented: the dialog continues from the point it was at before the interruption, regardless of the time that has passed. Similarly, when an initial topic of conversation gives way to a series of subtopics, so as to resemble spontaneous speech, movie dialog still tends to “stick to the point” (Taylor 1999:267), whereas in spontaneous spoken conversation totally different subjects can easily emerge (Taylor 1999) and then be abandoned.

Another artificial strategy which is usually used to help the audience keep track of the

movie is the introduction of a carefully planned rhythm of the dialog, which is slower and clearer than in naturally occurring conversation (Pavesi 2005:32). This is closely bound with another difference from spontaneous spoken conversation which is pointed out by Quaglio and Biber (2006: 716-717), in their comments on the TV series *Friends*: the TV series “has almost no overlaps, to avoid possibility of misunderstandings by the audience” and “at the discourse level, there are far fewer repetitions and interruptions” than those usually found in natural conversation (cf. also Pavesi 2005:32). In a similar way, features which usually give movie language a slant of spontaneity and are usually abundant in real conversation, like discourse markers (cf. Chaume 2004b:850; Pavesi 2005:32; Forchini *forthcoming*) and vocatives (Pavesi 2005:32), but which can be considered redundant in movies, do not show a very high frequency of occurrence.

As a non-spontaneous feature of movie dialog, Pavesi (2005:33) also points out the leveling out of sociolinguistic variation (i.e. dialectal traits, local and colloquial tones are often deleted and/or simplified), syntactic structures (i.e. monoclausal utterances are usually preferred and subordination tends to be distributed homogeneously, cf.; Pavesi 2005:32 and also Rossi 2003:103), lexical choices (i.e. usually movies offer the same core vocabulary, avoiding literary and dialectal terms, jargon and technicisms, cf. Pavesi 2005:33), turn taking and utterances (i.e. the latter tend to employ the same number of words, cf. Pavesi 2005:32) and dialogs (like, for example, the reduced and predictable use of phatic devices, interjections and discourse markers, Pavesi 2005:34), which are often stereotyped. Nevertheless, it is worth highlighting that some of these features, like the simplification of syntactic structures (cf. Biber *et al.* 1999) and the use of a core vocabulary (cf. McCarthy 1999:2; Chapter 2; and Section 3.2.2.1 here) are traits which can also be found in spoken conversation.

To sum up, if on the one hand, these distinctive characteristics of movie dialog largely contribute to its non-spontaneity, on the other, they are necessarily “imposed by the televized medium” (Quaglio and Biber 2006:716) and are needed to fulfill a number of functions, such as contributing to the unfolding and comprehension of the narrative. Clear and concise dialogs, together with explicit and linear breaks and blocks of information, help the audience understand what is going on (Taylor 1999, Pavesi 2005). It can be concluded, then, that the artificiality of movies is important. Interestingly, the fact that the audience seems to easily accept the non-spontaneous anomalies of movies (Pavesi 2005:30) proves that movie non-spontaneity is not limiting, but rather it is a useful device to get messages across. Besides, the

fact that movie dialogs imitate reality without including all the features which are typical of spontaneous spoken discourse (Taylor 1999, Chaume 2004b; Pavesi 2005) makes them a peculiar conversational domain.

3.2.2 Sounding Spontaneous: the Role of Language and the Movie Scriptwriter

Particularly in the past, the planned language of many movies was “stylized and patently false” and sounded absolutely artificial (Taylor 1999:264; cf. also Pavesi 2005:32). Taylor (1999:264) maintains that today the situation appears to be different, pointing out that, although many movie scriptwriters still produce inadequate scripts, they do produce convincing dialogic scenes, probably because they take into account Halliday’s (1985a) meta-functions of language (i.e. the ideational, interpersonal and textual one), which have particular relevance to understanding how the spoken language is used.

Since movie language is prefabricated and artificially designed to sound like authentic speech (Sinclair 2004b, Taylor 1999, Pavesi 2005), in order to sound authentic, it needs to be marked by features that are usually considered typical of spontaneous conversation. So, if the role of movie language is to “contribute to the evolution of the narrative, typify the characters and/or make them more realistic, and supply comments on the action” (Remael 2001:16), the role of the movie scriptwriter is to make all this happen. In order to do so, there are a number of factors, beyond the movie constraints mentioned above, that the movie scriptwriter has to take into account.

First and foremost, considering the ideational, interpersonal and textual meta-functions of language (Halliday 1985a) may help to plan language exchanges in a proper way, so that they can fit into the dramatic context without being too verbose or carrying obvious messages from the filmmaker to the viewer (Remael 2001). The ideational meta-function of language is an important component in that it is concerned with the expression of content, the information exchanged (Halliday 1985a); the interpersonal meta-function of language, instead, is relevant for it is linked to the use of language to establish and maintain social relations (Halliday 1985a). Finally, the textual meta-function of language deals with the way speakers organize their language into coherent text within the different contexts in which they interact (Halliday 1985a). Of course, knowing these meta-functions becomes extremely

relevant in contexts where the movie scriptwriter wants to simulate spontaneity: to be able to “reflect the predictable conversational patterns that are expected in real-life situations” (Taylor 1999:264), the scriptwriter needs to consider the content of real-life situations, the types of social talk and verbal interaction that his/her speakers use in conversational exchanges with family, friends, colleagues, etc. (Hawes and Thomas 1994:22), and the text and context in which speakers interact. In order to do so, when building up the movie dialog movie scriptwriters need to be familiar with some fundamental linguistic factors, of which Taylor (1999:264-265) lists a few:

- . whether the speakers know each other, or not, and if so, to what extent;
- . the speed at which speakers utter words, for some speakers are faster than others;
- . whether the speakers express themselves using complete clauses, or various degrees of ellipsis, for instance;
- . speakers who are used to initiating discourse use more *declaratives*; those who do not use them usually cover a secondary role; those in real or imagined positions of authority employ *imperatives*;
- . speakers use questions differently: information speakers use more *polar interrogatives*, whereas more conversational speakers prefer *wh-interrogatives*; women usually use *tag questions* especially when attempting to generate conversation;
- . some speakers use certain pronouns more than others: the constant or exclusive use of the *first personal pronoun* suggests egocentricity;
- . speakers use modality differently: a limited use of modals shows either assertiveness or unwillingness to express doubt, certainty, opinion, etc.

All these linguistic factors, which are, of course, not comprehensive, intertwine with the Hallidayan ideational, interpersonal and textual meta-functions of language, and are factors that the movie scriptwriter needs to consider when constructing the movie dialog if (s)he wants his/her script to sound spontaneous. Apart from these, the movie scriptwriter also needs to take account of sociolinguistic factors which deeply influence the way that speakers speak. In other words, scriptwriters have to examine linguistic structures within society as a whole, and concrete communicative situations (Thomas 1995, Berruto 2004) by spotting the proper diatypic (i.e. the register), diastratic (i.e. social class, age, sex, ...) and diatopic (e.g. region, country, ...) variety of his/her speaker(s). In much the same way, movie scriptwriters have to be

mindful of anthropological and psycholinguistic factors like the speaker(s)'s culture, mental status, mood, age, speech (dis)fluency, *inter alia*, which deeply influences speech processes, if the movie speakers are to sound real and authentic (Thomas 1995).

As illustrated by Bubel (2008), another factor which has to be taken into account when designing movie dialog is the role of the audience, in that the audience partakes in the co-construction of the meaning of the interaction. Bubel (2008) argues that the cognitive processes in 'screen-to-face' discourse are generally parallel to those of over-hearers in everyday situations:

Overhearers can only make conjectures about what they are able to listen in on, as they do not fully share the participants' common ground. Consequently, in order to be intelligible, film dialogue has to be carefully designed for overhearers so that they can reconstruct the participants' common ground, and the film production crew involved in this design has to construct the dialogue on the basis of the knowledge patterns they expect the future audience to share with them (Bubel 2008:69).

Bubel (2008) starts from Goffman's (1976, 1979) concept of listener roles, which divides them into overhearers, ratified participants, and addressees: the first type (overhearers) implies unratified participation, which can be either intentional or unintentional and can be either encouraged or discouraged; the second (ratified participants) type is associated to listeners who are not specifically addressed by the speaker, whereas the third type (addressees) refers to those listeners who are 'oriented to' by the speaker as if they were talked to directly. Overhearers can be further divided into bystanders and eavesdroppers; the former are not part of the conversation, but are openly present (e.g. "a couple is having a conversation on the bus, and people are sitting opposite them within hearing distance of what is said", Bubel 2008:61), while the latter "listen in without the speakers' being aware of it, for example, when someone is listening behind a door to a conversation going on inside a room" (Bubel 2008:61). Bubel (2008) maintains that the audience of the movie has a role similar to the over-hearers', in that they are unlikely to take part in all of the participants' shared experiences, and thus there is always some part of the common ground that is sealed off from them. Consequently, they can only draw inferences

about what speakers mean.

So as to explain how movie dialog has to be constructed for the implied audience, Bubel (2008) compares screen-to-face conversation to face-to-face conversation following Clark and Schaefer (1992), who posit the existence of four attitudes towards over-hearers: indifference, disclosure, concealment, and disuiselement. *Indifference* means that the speaker builds the utterance without paying any particular attention to the over-hearer; Bubel (2008:65), however, points out that “the speaker is still bound by politeness obligations; for example, when you are having a conversation on a bus, while sitting across from an elderly lady, you might refrain from using strong expletives and shouting”. *Disclosure* means that the speaker wants the over-hearer to acquire some information from the conversation, although he does not want him/her to actively take part in it (e.g. “the conversationalists on the bus vaguely know the person sitting opposite and, for example, want that person to gain a positive opinion of them” Bubel 2008:65). In order to do so, the speaker has to provide the over-hearers with enough evidence to easily draw correct inferences. *Concealment* implies that the speaker takes advantage of the lack of shared knowledge so that the over-hearer cannot gather information. Finally, *disuiselement* implies that the speaker wants the over-hearer to jump to the wrong conclusions without realizing it.

In movies, of course, the default attitude has to be disclosure, for it is fundamental that the audience understands what is going on. Consequently, utterances and turns need to be constructed to allow interpretations: “utterances are designed with overhearers in mind, on the basis of an estimate of the spectators’ world knowledge and on the knowledge the participants have gleaned from interactions that the spectators have observed” (Bubel 2008:66). In other words, the scriptwriter has to keep in mind “the disadvantages of screen-to-face discourse” (Bubel 2008:66) making relevant information explicit in the utterance. Bubel (2008:66) gives an example of two people in a science fiction movie “talking about a technological gadget that is standard equipment in their world but does not exist in the world of the audience” who “will have to include information on how it works and what its purpose is, even though both of them know” in order to “make this information available to the overhearer who does not share their common ground”. All this is hard to achieve, especially if the different background of movie audience is considered (i.e. movie viewers have different ages, genders, occupations, experiences, knowledge patterns, etc., Bubel 2008) and the question is undoubtedly further complicated by the fact that a script, which is planned to be

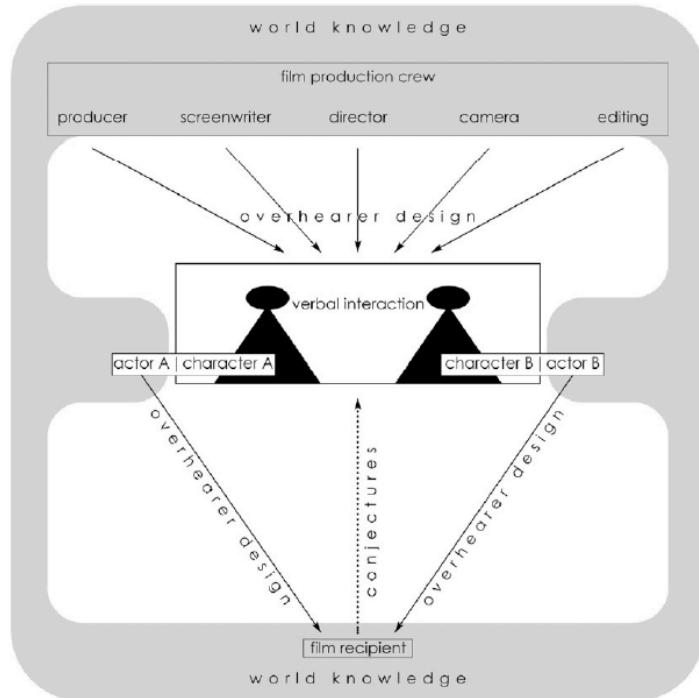
spoken by many speakers, is usually written and planned to sound authentic by a single scriptwriter (Taylor 1999, Pavesi 2005).

Undoubtedly the list of the features and factors illustrated is not at all exhaustive; indeed, it does not aim to be. It has been provided, first of all, to show that to plan a good movie script that sounds spontaneous, movie scriptwriters need to consider variables, which, added to the constraints of movie dialogs mentioned above (cf. Section 3.2.1.1), make total spontaneity rather impossible to achieve, or in Taylor's (1999:277) words, "restrict total authenticity":

The more successful attempts (Woody Allen, Spike Lee) would seem to be those where the writer is more immediately able to identify with his or her context, and when attention is paid to the interpersonal as well as the ideational content of the text. Cases where a higher percentage of *planning* is more evidently required (*Tootsie*, *Pulp Fiction*) produce highly successful films but more transparently constructed dialogue (Taylor 1999:277).

Second, the present listing has been provided to highlight that "the co-construction of meaning in movie discourse becomes a joint effort of the film recipients, the actors, the director, the screenwriter, the producer, the camera staff, and the cutters involved in the editing process" (Bubel 2008:68). This is illustrated by Figure 6, which shows that the members of both the production crew and of the audience make use of their world knowledge to design and interpret the movie dialog.

Figure 6. A model of movie discourse Bubel (2008:68)



3.2.2.1 Linguistic Features

Keeping in mind that the multiple codes involved in movies should be interpreted and considered as one whole multimodal code, since communication is delivered and compensated via their simultaneous interplay, the present section focuses specifically on the linguistic code. The aim is to highlight linguistic traits which are peculiar to movie language and help it by-pass its fictitious nature by gathering a touch of spontaneity.

As mentioned above, the peculiarity of movie language lies in the fact that it “purports to be authentic but conceals its hidden identity as planned discourse” (Taylor 1999:203). On the one hand, movie language is undoubtedly an artificial simulation, for it is prefabricated by birth, i.e. actors have to follow a script (Rossi 2003:94; Chaume 2004a:168) and act as if there were nothing to follow; on the other, taking into account the determinants of spontaneous spoken conversation identified in Chapter 3, at an abstract level it cannot be denied that non-spontaneous conversation shows similar traits to spontaneous conversation: (a) non-spontaneous conversation takes place in the spoken medium and occurs with non-verbal paralinguistic features; (b) non-spontaneous conversation pretends to take place in

real time but actually it takes place in real time, if “real time” is perceived as an ongoing process. Although movies are pre-recorded and not impromptu, the perception of the audience is that something is happening while watching the movie: “the visual medium with moving images and the potential of exploiting the written and spoken codes at the same time enhances the sense of immediacy” (Mansfield 2006:34; cf. also Pavesi 2005:30); (c) non-spontaneous conversation usually takes place in a shared context; (d) non-spontaneous conversation is interactive, continuous, and expressive of politeness, emotion, and attitude. At a more practical level, features (a) and (c) imply reliance on implicit meaning or reference and allow elaboration or specification of meaning avoidance as in spontaneous conversation. Consequently, deictics, such as *it, I, you, my*, (Taylor 1999:269; cf. also Rossi 2003:101) and elisions (Taylor 1999:275) are often inserted in movie dialog, making it sound more authentic. Feature (b) typically gives way to *normal dysfluency* and *fragmented language*, which, according to the literature (cf. Taylor 1999, Rossi 2003, Pavesi 2005), are spontaneous conversation traits that are also present in movies: speakers, for example, produce incomplete utterances (Rossi 2003:96), self-corrections/repairs (Rossi 2003:96,103), reformulations, repetitions (Rossi 2003:96,103), insert breaks/pauses (Rossi 2003:96), and/or overlapping conversation (Taylor 1999:273).

Finally, features (b) and (d) lead speakers to use the same repertoire of expressions like inserts (Quaglio and Biber 2006:716), hesitators (Quaglio and Biber 2006:716), vocatives (Taylor 1999:274; Quaglio and Biber 2006:716) hedges (Quaglio and Biber 2006:716) adjacency pairs (question/answer) (Taylor 1999:268), short and phatic devices (Taylor 1999:272), expletives (Taylor 1999:274), fillers (Taylor 1999:274), tag questions (Taylor 1999:276), and discourse markers (Taylor 1999:269; Rossi 2003:96; Quaglio and Biber 2006:716; Forchini *forthcoming*); all devices which help to keep the conversation going. Movie language vocabulary is another level which offers further evidence of speakers using repetitive structures when talking; indeed, it seems to favor a core vocabulary, which usually avoids literary and dialectal terms, jargon and technicisms (Rossi 2002:161, cf. also Rossi 1999; Pavesi 1994, 1996, 2000; Taylor 1999). This is a very interesting trait for it recalls the basic, or core, vocabulary typical of spontaneous spoken conversation (cf. McCarthy 1999:2 and Chapter 2).

Another feature of movie language which brings to mind spontaneous conversation is the use of different conversational environments (Pavesi 2005:30), like dialogic exchanges

between colleagues, friends, neighbors, and conversations in restaurants, at the mall, at the hairdressers', etc. (Pavesi 2005:30); the use of a-symmetric interactions like superior-inferior, doctor-patient, teacher-learner (Pavesi 2005:30); and the use of plurilinguism, code switching (i.e. the movement from one language to another), and code mixing (i.e. hybridization) (Rossi 2003:113).

All the features typical of spontaneous spoken conversation illustrated so far also add a tone of spontaneity to movie language in terms of language informality and interpersonality: they make spontaneous conversation more informal in that they show it is not influenced by the prestige, wellformedness and correctness typical of written texts (Biber *et al.* 1999:1050) and they highlight the interpersonal meta-function of language for they are employed to establish and maintain social relations (Halliday 1985a). The same can be said for the use of syntactic-pragmatic strategies like contractions (Quaglio and Biber 2006:716), fronting (Taylor 1999:275), dislocations (Taylor 1999:276; Pavesi 2005:22), clefts (Pavesi 2005:102), *inter alia*; and for the use of two-grams like *are you, do you, all right, come on, thank you*, etc. (Forchini *forthcoming*), which highlight the informal and interpersonal dialogic character which is typical of spontaneous conversation (Biber *et al.* 1999).

Table 12 below summarizes the features of movie conversation which are usually inserted to make it sound spontaneous.

Table 12. Spontaneous traits of movie conversation

SPONTANEOUS TRAITS OF MOVIE CONVERSATION				
MC takes place in the spoken medium				
MC occurs with non-verbal paralinguistic features	. implicit meaning and reference . no elaboration or specification of meaning	. deictics . elisions		
MC takes place in shared context				
MC pretends to take place in real time giving a sense of immediacy	. normal dysfluency . fragmented language	. incomplete utterances . self-corrections/repairs . reformulations . repetitions . breaks/pauses . overlaps	. informality . interpersonality	. contractions . fronting . dislocations . clefts . interpersonal two-grams
MC is interactive, continuous, expressive of politeness, emotion, and attitude	. same repertoire of expressions	. inserts . hesitators . vocatives . hedges . adjacency pairs . short and phatic devices . expletives . fillers . tag questions . discourse markers . core vocabulary		
. different conversational constellations . a-symmetric interactions . plurilinguism . code switching . code mixing				

3.2.2.2 Discourse Markers: the Case of *You Know*

Although movie dialogs are written to imitate real dialogs (Nencioni 1976, Taylor 1999, Rossi 2003, Chaume 2004b, Pavesi 2005), and discourse markers are highly frequent in spontaneous spoken conversation (Biber *et al.* 1999), Chaume (2004b:850) points out that they do not appear so frequently in movies:

Although film dialogues want to imitate real dialogues, it is striking that in a whole film *you know* only appears five times. Markers such as *you know* or *I mean* are abundant in real conversation. Film dialogues form part of what it is called prefabricated discourse: it imitates reality but cannot include all the hesitations, repetitions and syntactic anomalies that actual oral discourse contains.

A similar pattern is pointed out in Forchini (*forthcoming*):

you know and *I mean* do not seem to be very frequent in movies. In the AMC, *you know* occurs 89 times, but only 46 occurrences are DMs, and *I mean* occurs only 37 times (all of which are DMs); in terms of a percentage, this means 0,12% of *you knows* and 0,1% *I means* in the whole corpus.

However, it needs to be taken into account that the low occurrence of DMs in movies may be due to a number of reasons: first of all, to the small size of the corpus used for the investigation, namely 70,000 words (i.e. nearly 8 hours of movie speech)⁴⁷; second, to the fact that discourse markers simply occur less or do not occur) in movies; and third, to the fact that they do occur in movies, but not in all types; indeed, in the sample used by Forchini (*forthcoming*) there is only one comedy and it has the highest number of frequency of DMs. Consequently, these results cannot but be interpreted as preliminary. The same can be said about Chaume's (2004b) results in that they are based on the study of the movie *Pulp Fiction* only (cf. Chaume 2004b:843), which, also, according to Taylor (1999:277), seems to show a high percentage of planning, and rather transparently constructed dialogue. It is also worth noting that *Pulp Fiction* belongs to an earlier period, i.e. 1994, which means its planning may be more accurate, cf. Taylor 1999 on older movies being more accurate.

Another aspect worth of note is that despite the low percentage of the two discourse markers mentioned in the corpus, in terms of two-grams, their position is relatively high (i.e. *you know* ranks at position 14 and *I mean* at position 26, which is quite high in a corpus

⁴⁷ The American Movie Corpus used by Forchini (*forthcoming*) at the time of writing consisted of the following movies: *Mission: Impossible II* (J. Woo, 2000); *Shallow Hal* (B. and P. Farrelly, 2001), *Ocean's Eleven* (S. Soderbergh, 2001); and *The Matrix Reloaded* (A. and L. Wachowsky, 2003).

containing 3450 two-grams) and occurs together with other expressions like *are you*, *do you*, *all right*, *come on*, *thank you*, etc. (Forchini *forthcoming*), which are usually employed to highlight the interpersonal function typical of dialog (cf. Halliday 1994). Both Chaume's (2004b) and Forchini's (*forthcoming*) data, indeed, show that this is the main function of *you know*. More specifically, Chaume (2004b:850), quoting Schiffrin (1987), points out that *you know* is:

used to express shared knowledge between speaker and listener, or between speaker and the rest of the members of the same culture, that is, "general consensual truths (Schiffrin, 1987:274). [...] it has a clearly interactional function expressing confidentiality between the speakers, a device used to bring the listener to your own field. This is why it is usually employed in through-arguments, "y'know appeals to shared knowledge as a way of converting an opponent to one's own side in a dispute" (Schiffrin, 1987:279).

Similarly, Forchini (*forthcoming*) maintains that *you know* is used especially under two circumstances: "either to guide the listener in the interpretation of the utterance or to allow the speaker time to find appropriate words". In particular, when used to guide the listener in the interpretation of the utterance, *you know* acquires the following functions:

. *telling/commenting function*, when *you know* is employed to add information, introduce a new topic or comment, as in example A:

A. Yeah, sure, Link. Hey, **you know**, *next year I'm old enough to join a crew*, right. I've been thinking a lot about it and I've made my decision. (The Matrix Reloaded)

. *clarifying function*, when *you know* is used to clarify/explain something, as in example B:

B. I need a man around that can *give it to me straight*, **you know**, *whether the news be good or bad*. So I've decided - from now on, you'll be working directly for me. (Shallow Hal)

. *knowledge marker*, when *you know* is used to appeal to knowledge or to awaken knowledge, as

in example C:

- C. Hal, we gotta go to do *that* thing. **You know**, at the at the place.
(Shallow Hal)

When, instead, *you know* is employed to allow the speaker time to find the most appropriate expression, the DM may be used either as a strategy to play for time, i.e. as a time staller, or “to fill the gap while the speaker seeks the right words”; or “as a mitigator/hedge to play for time to find the right words, because the speaker does not know what to say or needs to gain time to soften the severity of the situation”, as in example D where Rosemary is not used to receiving compliments or being asked for date and before saying so she employs a DM (i.e. *you know*), a vocative (*Hal*) and a syntactic blend (i.e. It’s just, **you know**, Hal, *I’m not...*):

- D. Speaker1: What? I thought we were having a good time.
Speaker2: We were. It’s just, **you know**, Hal, *I’m not used to all this*.
(Shallow Hal).

Forchini (*forthcoming*) highlights the fact that in spite of its multi-functionality, *you know* mostly occurs with the *telling function*. This recurrent co-occurrence suggests that the *telling function* is a key function to the pragmatics of *you know*, and if so, the pragmatic meaning of DM is closer to its literal meaning of *you know*, rather than to its non-discourse-marker-like meaning (cf. also Schiffarin 1987).

Regarding functions in context, Forchini (*forthcoming*) provides a description of the functions that *you know* acquires according to its turn position: as illustrated by Table 13, *you know* occurs with the highest frequency in mid position (i.e. 52.17%), but also has a rather high percentage of occurrence in initial position (i.e. 32.6%). Interestingly, when *you know* occurs in initial and mid-position, it is especially used with a telling function, whereas in final position it does not show any preference. However, in spite of the highest percentage in mid-position, *you know* may occur anywhere with any function (except in initial position, where it is not used as a shared knowledge marker).

Table 13. *You know* Functions and Utterance Position in the AMC (Forchini *Forthcoming*)

YOU KNOW FUNCTIONS & UTTERANCE POSITIONS in the AMC				
FUNCTION	INITIAL	MEDIAL	FINAL	TOTAL
TELLING FUNCTION	9	11	2	22
CLARIFYING FUNCTION	4	7	1	12
SHARED KNOWLEDGE MARKER		4	2	6
TIME STALLER	2	2	2	6
TOTAL	15	24	7	46

So, if on the one hand, the low occurrence of *you know* in movies makes it part of the fictitious side of movie dialog, on the other, its pragmatic function recalls the interpersonal function typical of spontaneous spoken conversation. Besides, “the most frequent functions, the highest occurrence in utterance mid-position, and the functions linked to utterance positions appear to be the same in the two conversational domains” (Forchini *forthcoming*).

CHAPTER 4. MULTI-DIMENSIONAL ANALYSIS

As pointed out in Chapter 3, movie conversation is usually considered by the literature as a kind of artificial and non-spontaneous speech, designed to sound like authentic language (Taylor 1999, Rossi 2003, Pavesi 2005). For this reason, it is often claimed that it does not truly reflect natural conversation and that it is “not likely to be representative of the general usage of conversation” (Sinclair 2004b:80).

Contrary to what is usually maintained, however, the findings listed in Table 14⁴⁸ from the LSAC and the AMC retrieved with the *Biber tagger* and the *SAS package*, reveal that face-to-face and movie conversation, in fact, do not differ much in that they share nearly all the linguistic features examined. The most frequent features in both domains, for example, are *verbs* (uninflected present, imperative and third person – *pres* in the tables), *second person pronouns* and *possessives* (*pro2*), *first person pronouns* and *possessives* (*pro1*), *nouns* (*n*), and *prepositions* (*prep*). The least frequent features are *wh pronouns* functioning as relative clauses in object position (*rel_obj*), *wh pronouns* functioning as relative clauses in subject position (*rel_subj*), *wh pronouns* functioning as relative clauses in object position with prepositional fronting (*rel_pipe*), *suaive verbs* (e.g. ask, command, insist – *sua_vb*), *passive verbs + by* (*by_pasv*), and *passive postnominal modifiers* (*whiz_vbn*).

⁴⁸ The *variables* in the table(s) are the linguistic features analyzed (for the meaning of the codes of the specific features see Appendix 1); *N* stands for the number of texts selected (with regard to movie conversation the number 3 refers to the three sub-genres, or sub-corpora, labeled comedies, border-line movies, and non-comedies); *Mean* is the mean (average) frequency of items. The frequency counts of all linguistic features are normalized to a text length of 1,000 words.

Table 14. Linguistic features of movie and face-to-face conversation

D I M E N S I O N S	Movie Conversation Linguistic Features			Face-to-Face Conversation Linguistic Features		
	Variable	N	Mean	Variable	N	Mean
	ffffffffff	ffffffffff	ffffffffff	ffffffffff	ffffffffff	ffffffff
1=	typetokn	3	53.5333333	typetokn	327	46.4253823
2=	wrdlngth	3	3.8333333	wrdlngth	327	3.6941896
3=	wordcnt	3	34622.67	wordcnt	327	5729.45
4=	ttnum	3	24.0000000	ttnum	327	29.0489297
{Dimension 1}						
1=	prv_vb	3	24.4000000	prv_vb	327	29.4944954
2=	that_del	3	8.5666667	that_del	327	9.8645260
3=	contrac	3	6.5666667	contrac	327	2.4464832
4=	pres	3	117.2333333	pres	327	118.2146789
5=	pro2	3	53.3666667	pro2	327	35.3773700
6=	pro_do	3	4.3666667	pro_do	327	3.2461774
7=	pdem	3	12.6000000	pdem	327	13.1455657
8=	gen_emph	3	9.1000000	gen_emph	327	11.8308869
9=	prol	3	72.3333333	prol	327	65.8051988
10=	it	3	19.0000000	it	327	24.6094801
11=	be_state	3	3.2000000	be_state	327	3.2195719
12=	sub_cos	3	1.4666667	sub_cos	327	2.3030581
13=	prtcle	3	7.7333333	prtcle	327	14.0073394
14=	pany	3	8.1000000	pany	327	7.8688073
15=	gen_hdg	3	1.5333333	gen_hdg	327	2.5079511
16=	amplifr	3	2.3333333	amplifr	327	2.3969419
17=	wh ques	3	5.6666667	wh ques	327	3.1651376
18=	pos_mod	3	8.4333333	pos_mod	327	8.3556575
19=	o_and	3	11.5666667	o_and	327	8.5972477
20=	wh_cl	3	2.4666667	wh_cl	327	2.5715596
21=	finlprep	3	3.9333333	finlprep	327	3.3131498
22=	n	3	191.4666667	n	327	186.4244648
23=	prep	3	63.4666667	prep	327	63.7510703
24=	adj_attr	3	16.3000000	adj_attr	327	17.5605505
{Dimension 2}						
25=	pasttnse	3	31.4666667	pasttnse	327	38.4792049
26=	pro3	3	24.2000000	pro3	327	31.9446483
27=	perfects	3	8.8000000	perfects	327	5.0917431
28=	pub_vb	3	5.2000000	pub_vb	327	6.5678899
{Dimension 3}						
29=	rel_obj	3	0.4666667	rel_obj	327	0.3299694
30=	rel_subj	3	0.6666667	rel_subj	327	0.5926606
31=	rel_pipe	3	0.1333333	rel_pipe	327	0.0730887
32=	p_and	3	0.6333333	p_and	327	0.9067278
33=	n_nom	3	13.0333333	n_nom	327	10.6296636
34=	tm_adv	3	6.6333333	tm_adv	327	8.0489297
35=	pl_adv	3	13.6333333	pl_adv	327	13.9813456
36=	advs	3	46.3666667	advs	327	56.9284404
{Dimension 4}						
37=	inf	3	12.3666667	inf	327	6.8620795
38=	prd_mod	3	7.9333333	prd_mod	327	7.0752294
39=	sua_vb	3	0.8333333	sua_vb	327	0.2529052
40=	sub_cnd	3	3.9000000	sub_cnd	327	4.5149847
41=	nec_mod	3	5.1333333	nec_mod	327	5.0370031
42=	spl_aux	3	2.2333333	spl_aux	327	2.4675841
{Dimension 5}						
43=	conjnts	3	6.8333333	conjnts	327	1.3749235
44=	agls_psv	3	4.6333333	agls_psv	327	3.0431193
45=	by_pasv	3	0.2000000	by_pasv	327	0.1525994
46=	whiz_vbn	3	0.5333333	whiz_vbn	327	0.4477064
47=	sub_othr	3	5.1666667	sub_othr	327	7.9715596

This preliminary observation is extremely important as it presupposes that the features that face-to-face and movie conversation share serve similar functions (cf. Biber 1988:91) and are evidence of similar textual Dimensions (cf. Biber, Conrad and Reppen 1998). If this is the case, namely, if empirical data show that movie language shares the same linguistic features with similar functions as natural conversation, the current view which considers movie dialog non-representative of the general usage of conversation will have to be re-considered.

Section 4.1 investigates this similarity through Multi-Dimensional techniques and provides a macro-overview of the general linguistic features present in the two conversational domains: Sections 4.1.1, 4.1.2, 4.1.3, 4.1.4, and 4.1.4 concentrate on Dimensions 1, 2, 3, 4, and 5 respectively. Section 4.2, instead, focuses on the comedy vs. non-comedy distinction in order to see whether movie genre influences this resemblance. Finally, Section 4.3 summarizes the Multi-Dimensional results.

4.1 Face-to-Face and Movie Conversation Compared

Multi-Dimensional analysis is used here to determine, in general, the co-occurrence of linguistic features, and, in particular, to verify the extent to which face-to-face and movie conversation differ or resemble each other. The assumption behind Multi-Dimensional analysis is that co-occurring linguistic features in a corpus (or in more corpora) share at least one communicative function, and that, by underlying each set of co-occurring linguistic features, it is possible to identify unified Dimensions (cf. Biber 1988:79 and Chapter 1 for further details).

Table 16 presents Multi-Dimensional data: the label *Variable* stands for the 5 Dimensions (or Factors, i.e. *dim1-5* in the table) taken into account; *N* for the number of texts (or sub-corpora) that made up the two corpora considered; *Mean* for the mean (average) frequency of items (the higher it is, the more frequent the items are); *Std Dev* for standard deviation, namely, a measure of the spread of the distribution⁴⁹; and *Minimum* and *Maximum*

⁴⁹ Biber, Conrad and Reppen (1998:280) explain that in all Multi-Dimensional studies “frequencies are standardized to a mean of 0.0 and a standard deviation of 1.0 before factor scores are computed. This process translates the scores for all features to scales representing standard deviation units, thus, regardless of whether a feature is extremely rare or extremely common in absolute terms, a standard score of +1 represents one standard deviation unit above the mean score for the feature in question. That is, standardized scores measure whether a

for the minimum and maximum frequencies of items respectively. In the characterization of the *Dimensions* (i.e. the *Factors* investigated), some linguistic features have negative weights and others positive weights along a polar continuum: Factor 1 (dim1) displays *informational versus involved production*, namely, a Dimension which marks “high informational density and exact informational content versus affective, interactional, and generalized content” (Biber 1988:107). Factor 2 (dim2) represents *narrative versus non-narrative concerns*, a Dimension which “can be considered as distinguishing narrative discourse from other type of discourse” (Biber 1988:109). Factor 3 (dim3) concerns *explicit versus situation-dependent reference*, a Dimension which distinguishes “between highly explicit, context-independent reference and nonspecific, situation-dependent reference” (Biber 1988:110). Factor 4 (dim4), the only Dimension which is only positive, is about *overt expression of persuasion*, a Dimension which “marks the degree to which persuasion is marked overtly” (Biber 1988:111). Finally, Factor 5 (dim5) reflects *abstract versus non-abstract information*, a Dimension which “seems to mark informational discourse that is abstract, technical, and formal versus other types of discourse” (Biber 1988:113). All Dimensions were further illustrated in Chapter 1 (cf. Section 1.3.1).

Table 15. Multi-Dimensional analysis of face-to-face and movie conversation

MULTI DIMENSIONAL ANALYSIS					
American Movie Corpus					
Variable	N	Mean	Std Dev	Minimum	Maximum
dim1	3	35.3166667	1.7013622	33.4100000	36.6800000
dim2	3	-0.9700000	0.2778489	-1.1500000	-0.6500000
dim3	3	-5.7233333	0.1887679	-5.9300000	-5.5600000
dim4	3	0.6466667	0.8195324	0.1100000	1.5900000
dim5	3	1.6633333	0.7011657	0.8700000	2.2000000
American Face-to-Face Conversation Corpus					
Variable	N	Mean	Std Dev	Minimum	Maximum
dim1	327	35.0451070	7.0665176	9.9800000	53.5800000
dim2	327	-0.8459327	1.3330098	-5.1000000	3.7900000
dim3	327	-7.0434557	2.0445282	-14.8100000	-1.1300000
dim4	327	0.6002141	2.0707167	-6.6100000	8.3400000
dim5	327	-2.0426911	0.7711589	-3.6300000	3.5000000

feature is common or rare in a text relative to the overall average occurrence of that feature. The raw frequencies are transformed to standard scores so that all features on a factor will have equivalent weights in the computation of Dimension scores. If this process was not followed, extremely common features would have much greater influence than rare features on the Dimension scores.”

In detail, Table 15 demonstrates that face-to-face and movie conversation have four Dimensions out of five in common: they both have a positive score with respect to Dimension 1 and 4 (which correspond to Factor 1 and Factor 4, namely, *informational versus involved production* and *overt expression of persuasion* respectively) and a negative score with respect to Dimension 2 and 3 (which correspond to Factor 2 and Factor 3, namely, *narrative versus non-narrative concerns* and *explicit versus situation-dependent reference* respectively). The only Dimension they differ in is Dimension 5 (which corresponds to Factor 5, namely, *abstract versus non-abstract information*). Dimensions 1, 2, 3 and 4, in particular, reveal that traits of spontaneity are present in both the conversational domains analyzed: informality, non-narrative concerns, situation-dependent factors, and a low level of persuasion are, indeed, typical features of spontaneous conversation⁵⁰ (cf. Chapter 2).

4.1.1 Dimension 1: Informational vs. Involved Production

The linguistic features which characterize the Dimensions and have negative and positive weights⁵¹ can be understood in terms of low and high frequency. In Dimension 1, for example, which reflects *informational versus involved production* (cf. Biber 1988), the following linguistic features have a negative weight: nouns, prepositional phrases, attributive adjectives, word length, and type-token ratio (respectively *n*, *prep*, *adj_attr*, *wrdlength*, and *typetokn* in the tables). This means that, if they are frequent, the production is more informational than involved: high frequency of nouns, the main bearers of referential meaning, is a sign of high density of information⁵²; prepositional phrases and attributive adjectives integrate information in a text; word length marks high density of information, for longer words convey more specialized meaning than shorter words; and type-token ratio depends on the use of many different lexical items in a text - this variation in vocabulary reflects an extensive use of words that have very specific meanings (cf. Biber 1988:104-105). The texts containing a high

⁵⁰ As illustrated in Chapter 2, in terms of Biber's (1988) Dimensions, spoken language is considered to be involved (Dimension 1), with non-narrative concern (Dimension 2), situation-dependent (Dimension 3), not particularly persuasive (Dimension 4), and non-abstract (Dimension 5).

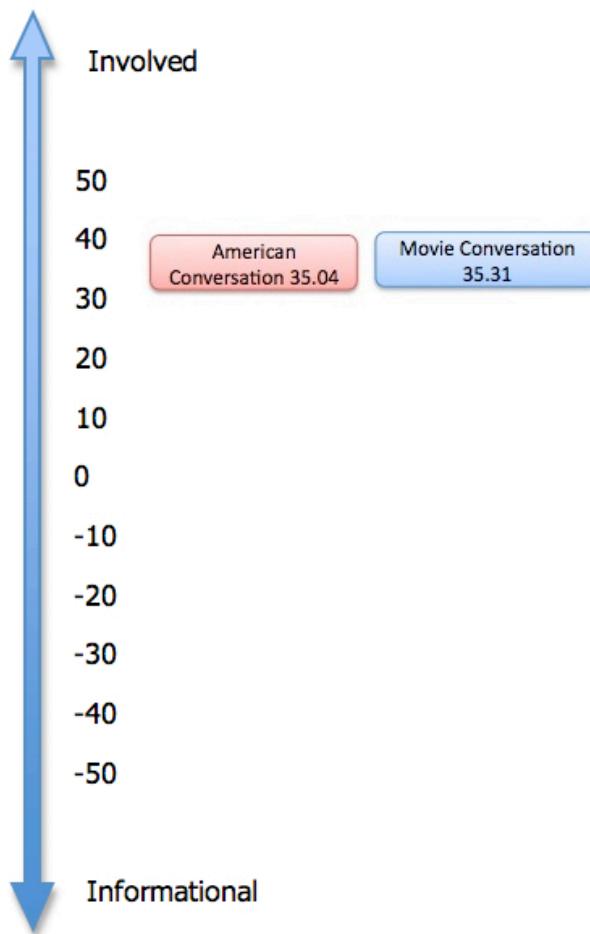
⁵¹ The features characterized by negative weights co-occur, as do those with positive weights (Biber 1988).

⁵² It is worth noting that, although the occurrence of nouns is high here, it reflects the usual occurrence in spoken language (cf. face-to-face conversation = 137.4 in Biber 1988:264) and it is relatively low compared to their frequency in written registers (cf. press reportage 220.5; press editorials = 201.0; press reviews = 208.3; official documents = 206.5; academic prose = 188.1; general fiction = 160.7; in Biber 1988:247-269).

number of occurrences of these linguistic features, then, are characterized by high informational content, since they present information as concisely and precisely as possible, similarly to written texts (cf. Biber, Conrad and Reppen 1998).

As Table 15 and Figure 7 illustrate, instead, both face-to-face and movie conversation are characterized by a positive mean score of 35.31 and 35.04, respectively, and thus their production is involved, rather than informational.

Figure 7. Dimension 1: *informational versus involved production*



This means that both face-to-face and movie conversation present a rather high affective, interactional, and generalized context typical of spoken language and distinctive of an interpersonal dialogic character (cf. Biber 1988). This similarity depends on the number and type of linguistic features that the two conversational domains share. As Table 16 and 17 illustrate, indeed, the frequency of those items which have a positive weight on Dimension 1 is

higher than the one of those which have a negative weight on it.

Table 16. Linguistic features of Dimension 1 of the LSAC

DIMENSION 1

means for conv 17:13 Monday, April 21, 2008 51

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
prv_vb	327	29.4944954	6.2505288	5.6000000	52.2000000
that_del	327	9.8645260	3.3464515	0	23.4000000
contrac	327	2.4464832	1.9780339	0	17.1000000
pres	327	118.2146789	13.8878361	66.5000000	166.7000000
pro2	327	35.3773700	9.3256379	8.4000000	92.4000000
pro_do	327	3.2461774	1.8262677	0	15.0000000
pdem	327	13.1455657	4.2197647	0	29.5000000
gen_emph	327	11.8308869	3.9893793	0	29.3000000
prol	327	65.8051988	13.3580473	25.1000000	100.4000000
it	327	24.6094801	5.8861178	9.2000000	49.7000000
be_state	327	3.2195719	1.6509476	0	11.3000000
sub_cos	327	2.3030581	1.3674760	0	8.0000000
prtcle	327	14.0073394	6.0216894	0	34.8000000
pany	327	7.8688073	2.3328554	0	15.0000000
gen_hdg	327	2.5079511	1.4234876	0	11.0000000
amplifr	327	2.3969419	1.6239739	0	11.1000000
wh ques	327	3.1651376	1.9488258	0	11.0000000
pos_mod	327	8.3556575	2.8813910	0	24.9000000
o_and	327	8.5972477	3.9154435	0	19.9000000
wh_cl	327	2.5715596	1.5464242	0	19.2000000
finlpref	327	3.3131498	1.6187300	0	14.9000000
typetokn	327	46.4253823	4.8230528	9.3000000	56.0000000
wrdlngh	327	3.6941896	0.1260412	3.2000000	4.3000000
n	327	186.4244648	29.7787941	120.6000000	326.2000000
prep	327	63.7510703	10.1261063	0	97.9000000
adj_attr	327	17.5605505	5.2182728	0	42.3000000

These features, consequently, contribute to affective, interactional, and generalized context: as illustrated briefly above, face-to-face conversation presents the highest mean score in *verbs* (uninflected present, imperative and third person – i.e. *pres* in the tables) and *second person pronouns* and *possessives* (*pro2*), namely, 118.21 and 65 respectively; and a relatively high mean score of *first person pronouns* and *possessives* (*pro1*), of *private verbs* (e.g. believe, feel, think – *prv_vb*), of *it pronouns* (*it*) and of *discourse particles* (e.g. now – *prtcle*), i.e. 35.37, 29.49, 24.60, and 14.00 in the order mentioned. All these items are associated to an *involved* Factor in that they contribute to a context that can be described as oral, affective, fragmented, interactional, and generalized⁵³: *private verbs*, for example, are used to express private attitudes, emotions and thoughts; *present tense forms* are employed to indicate actions taking place in the immediate context of the action; *first* and *second person pronouns* are highly present in interactive discourse; *it* (together with *demonstrative* and *indefinite pronouns*)

⁵³ *Private verbs* and *present tense forms* are the features bearing largest weight on this Dimension for they are indicators of a verbal style, as opposed to nominal style (cf. Biber 1988:105). The other terms which have positive weight on Factor one are listed in Table 16 and in Appendix 1.

stands for unspecified nominal referents; and *discourse particles* are generalized markers of information which help to maintain textual coherence (cf. Biber 1988:104-108).

The following extract from the LSAC shows the high frequency of these features in spoken conversation (examples in bold):

Extract 3. Features characterizing Dimension 1, from the LSAC

Speaker1: Did you manage?

Speaker2: **Yeah.**

Speaker1: **Well**, how, **that's** very clever of **you**. I've been trying to open one for.

Speaker2: Do **you** have fingernails?

Speaker1: **Yeah** **you** have fingernails. **You** should be able to get that one?

Speaker2: **Oh** this **is** great.

Speaker1: **Yeah, yeah**, except they are not really very good quality. But at least they are very small [so you]

Speaker2: [Yeah].

Speaker1: **You** don't need much space for that.

Speaker2: **Um**, ... maybe I'll let **you** tell **me** how, how **it opens up**.

Speaker1: **Uh, I think it's just ... pulling here right?**

Speaker2: **Uh huh.**

Speaker1: And then **I guess you need** to just take ... take them, **I** don't know if **you** want to take them all out or just leave them like that. There's a lot of <unclear>.

(American face-to-face conversation)

Similarly, as Table 17 shows, movie dialogs present a high percentage (i.e. more than 50%) of *verbs*, and *first* and *second person pronouns/possessives*, which give a rather affective, interactional, and generalized context to this type of domain.

Table 17. Linguistic features of Dimension 1 of the AMC

DIMENSION 1

means per 1,000 words for movies combined = 3 together						48
						17:13 Monday, April 21, 2008
The MEANS Procedure						
Variable	N	Mean	Std Dev	Minimum	Maximum	
<i>ffffffffff</i>						
POSITIVE						
prv_vb	3	24.400000	0.7211103	23.800000	25.200000	
that_del	3	8.5666667	0.2309401	8.300000	8.700000	
contrac	3	6.5666667	1.3650397	5.000000	7.500000	
pres	3	117.2333333	2.9022979	114.700000	120.400000	
pro2	3	53.3666667	1.0969655	52.500000	54.600000	
pro_do	3	4.3666667	0.3785939	4.100000	4.800000	
pdem	3	12.600000	0.7810250	11.700000	13.100000	
gen_emph	3	9.100000	2.1702534	7.700000	11.600000	
prol	3	72.3333333	2.8501462	69.500000	75.200000	
it	3	19.000000	1.5000000	17.500000	20.500000	
be_state	3	3.200000	0.500000	2.700000	3.700000	
sub_cos	3	1.4666667	0.2309401	1.200000	1.600000	
prtcle	3	7.7333333	0.9504385	6.800000	8.700000	
pany	3	8.100000	1.2489996	7.100000	9.500000	
gen_hdg	3	1.5333333	0.4725816	1.000000	1.900000	
ampliffr	3	2.3333333	0.3511885	2.000000	2.700000	
wh_ques	3	5.6666667	0.7023769	5.000000	6.400000	
pos_mod	3	8.4333333	1.8147543	7.100000	10.500000	
o_and	3	11.5666667	0.4618802	11.300000	12.100000	
wh_cl	3	2.4666667	0.1154701	2.400000	2.600000	
finlprep	3	3.9333333	0.6429101	3.200000	4.400000	
NEGATIVE						
typetokn	3	53.5333333	4.5456939	48.300000	56.500000	
wrdlngth	3	3.8333333	0.0577350	3.800000	3.900000	
n	3	191.4666667	2.5658007	189.300000	194.300000	
prep	3	63.4666667	3.2254199	60.500000	66.900000	
adj_attr	3	16.300000	1.5874508	15.100000	18.100000	

The following passage from the AMC provides examples (in bold) of the linguistic features that characterize Dimension 1:

Extract 4. Features characterizing Dimension 1, from the AMC

Speaker1: Hey Russ! Rusty. What's up man?

Let me ask you a question now. Are you incorporated?

Roll, okay, if you are not, you should really think about it, cos I was talking to my manager last night...

Speaker2: Bernie?

Speaker1: No, not Bernie my business manager. Actually.

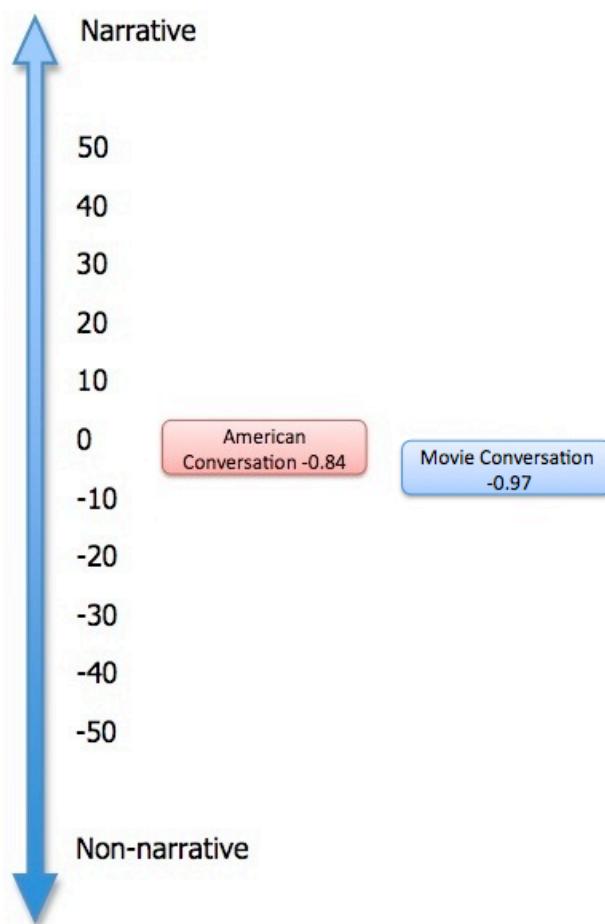
You know they're both named Bernie. Anyway, he was telling me that because of what we do, can be considered like research. For like a future. Gig or whatever. I can totally make it a tax write-off, the one thing is and this is, like, just his thing, and it's stupid. But. I'd have to pay you by check.. What?. Let's, or we could just stick to cash. Yeah, let's..., yeah, let's just stick to cash.

(American movie conversation)

4.1.2 Dimension 2: Narrative vs. Non-Narrative Concerns

Face-to-face and movie conversation also display extremely similar variables on Dimension 2 (*narrative versus non-narrative concerns*, cf. Biber 1988): they both have a negative score (face-to-face conversation has -0.84 and movie dialog has -0.97, cf. Figure 8), which means that they are both characterized by non-narrative concerns and are, thus, marked by immediate time and attributive nominal elaboration.

Figure 8. Dimension 2: *narrative versus non-narrative concerns*



In linguistic terms, this depends on the fact that face-to-face conversation (cf. Table 18) is characterized by a low occurrence of *verbs in the perfect aspect* (*perfects* in the tables), *public verbs* (e.g. assert, complain, say, report, declare – *pub_vb*), of *past tense verbs* (*pasttense*), and of *third person pronouns* except *it* (*pro3*), which are all devices that mark narrative discourse

and have a positive weight on Dimension 2 (cf. Biber 1988:109). Indeed, *past tense* and *perfect aspect* mark past events; *public verbs* are used to indicate indirect, reported speech; *third person pronouns* (except *it*) are used to refer to specific animate referents described in the narrative discourse (cf. Biber 1988:109).

Table 18. Linguistic features of Dimension 2 of the LSAC

DIMENSION 2

means for conv	17:13 Monday, April 21, 2008 51				
The MEANS Procedure					
Variable	N	Mean	Std Dev	Minimum	Maximum
ffffffffff	327	38.4792049	11.5915373	0	72.2000000
pasttnse	327	31.9446483	12.7553704	0	96.9000000
pro3	327	5.0917431	2.0774776	0	14.9000000
perfects	327	6.5678899	3.1678082	0	23.4000000
pub_vb	327				

The following extract – which is a randomly chosen example from the LSAC – shows, indeed, that there are no, or very few, perfects, past tenses, third person pronouns except *it*; on the other hand, there is the present tense which, together with attributive adjectives, has a negative weight on this Dimension (some examples in bold). These elements, which have a positive on Factor 1 (see above), have, instead, a negative score on Factor 2⁵⁴; this explains why face-to-face conversation has a negative (i.e. non-narrative) Dimension 2.

Extract 5. Features characterizing Dimension 2, from the LSAC

Speaker1: **It's** a basement.

Speaker2: **Half** windows?

Speaker1: Yeah.

Speaker2: You **can visit**?

Speaker1: Absolutely. I've **got** plenty of space now, boy. And a Jeep.

Speaker2: <unclear>

Speaker1: Madge. Oh she will be in her glory.

Speaker2: **It's** great.

Speaker1: She won't have to go in the elevator.

Speaker2: She'll be able to get outside and outdoors.

⁵⁴ Biber (1988:109) explains this by highlighting that “a discourse typically reports events in the past or deals with more immediate matters, but does not mix the two”.

Speaker1: How long until you will be living in **it**?

Speaker2: I'm gonna move in, in a week or so after I get back.

Speaker1: What **is** the room that has the wood paneling on the walls?

Speaker2: That's the basement.

Speaker1: Oh. **It's got** windows.

Speaker2: Yeah. **It's** actually kind of raised. The whole thing **is** raised up so it's

Speaker1: Yeah I see steps going up.

Does it rain a lot?

Speaker2: **It rains** a lot in Chicago but the water, the basement **doesn't** **get** any water.

Speaker1: **I'm** sure he looked into that.

Speaker2: That was one of the things you **have to** check for.

<unclear> Santa Barbara.

Speaker1: Oh yes.

Yeah I just found out that a friend of mine is going to the University of Chicago to get her Ph D. I really **want** to go visit her. Maybe I'll come out and <unclear>.

Speaker2: <unclear>

Speaker1: Oh **is** she?

Speaker2: Yeah.

Speaker1: Oh good.

Speaker2: <unclear>

Speaker1: I understand <unclear> gonna be in nineteen ninety-four.

Speaker2: I hope we won't get any **student** loans after ninety-six.

Speaker1: <unclear> stretch **it** out.

Speaker2: I won't be able to <unclear> my student loans after ninety-six.

<unclear>

Speaker2: Push his arm.

Speaker1: Huh?

Speaker2: Push his arm.

Speaker1: Yeah.

Speaker2: But **it's** fun anyway.

Speaker2: **That's what's** important. If you like **it**.

Speaker1: Yeah.

In much the same way, movie dialogs (cf. Table 19) are characterized by very few occurrences of *past tense verbs*, *third person pronouns*, *verbs in the perfect aspect* and *public verbs* (e.g. *assert*, *complain*, *say*).

Table 19. Linguistic features of Dimension 2 of the AMC

DIMENSION 2

means per 1,000 words for movies combined = 3 together						17:13 Monday, April 21, 2008
The MEANS Procedure						
Variable	N	Mean	Std Dev	Minimum	Maximum	
ffffffffff	3	31.466667	2.5006666	28.600000	33.200000	
pasttense	3	31.466667	2.5006666	28.600000	33.200000	
pro3	3	24.200000	0.6557439	23.500000	24.800000	
perfects	3	8.800000	0.3605551	8.400000	9.100000	
pub_vb	3	5.200000	1.1532563	4.300000	6.500000	

The following example from the AMC well illustrates the absence of these features and highlights those which have a negative score on Dimension 2:

Extract 6. Features characterizing Dimension 2, from the AMC

Speaker1: Mrs Larson? **I**t uh **i**t won't be much longer, Mrs Larson.

Speaker2: Oh well **is** he in a lot of pain?

Speaker1: No No no. There will be no more pain for your husband **H**e's heavily sedated.

Speaker2: OK I **think** I'm gonna go, send **little** Hal in now.

Speaker1: No No no I **don't think that's** such a **good** idea. With all the painkillers uh the reverend's not exactly himself.

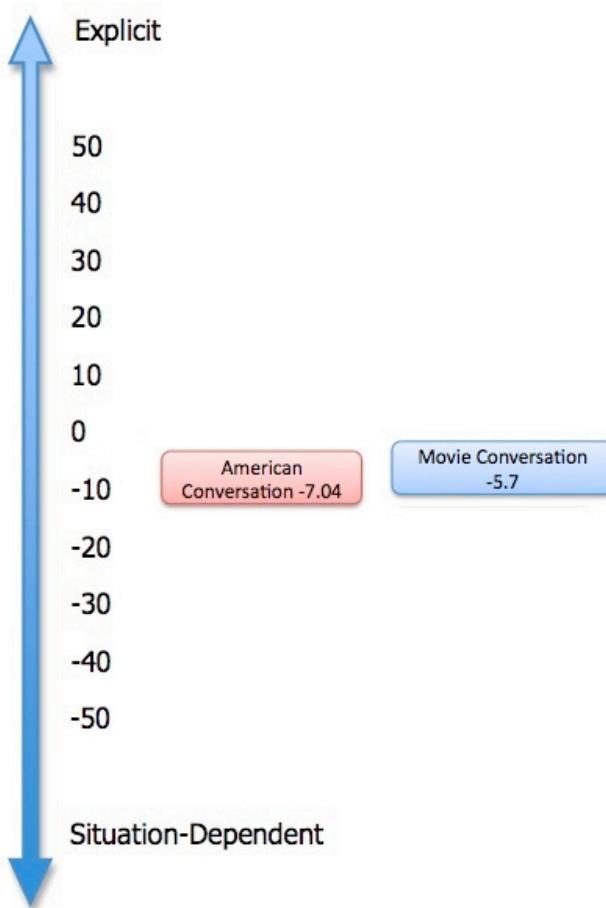
Speaker2: Look I **think** my boy **has** a right to say goodbye to his father I **mean** the man **means** everything in the world to him.

4.1.3 Dimension 3: Explicit vs. Situation-Dependent Reference

Face-to-face and movie conversation display extremely similar linguistic variables also with regard to Dimension 3 (*explicit versus situation-dependent reference*, cf. Biber 1988): they both

have a negative mean score (-7.04 and -5.7 respectively, Figure 9), which implies that they both rely on situation-dependent reference (cf. Biber 1988).

Figure 9. Dimension 3: *explicit versus situation-dependent reference*



More specifically, Multi-Dimensional analysis reveals (cf. Table 20) that face-to-face conversation has an extremely low mean score of *wh pronouns* that function as a relative clause in object position (*rel_obj* in the tables); *wh pronouns* that function as a relative clause in subject position (*rel_subj*); *wh pronouns* that function as a relative clause in object position with prepositional fronting (*rel_pipe*). All of these pronouns, indeed, are usually used as devices for the “explicit, elaborated indication of referents in a text” (cf. Biber 1988:110). Face-to-face conversation also displays a low percentage of phrasal connectors (*p_and*), nominalization (*n_nom*), which indicate referential and informational discourse. All these items have a positive weight on Dimension 3; lacking them implies having a negative Dimension and, consequently, being situation-dependent. Conversely, the highest occurrence

regards those items such as place and time adverbs (*pl_adv* and *tm_adv*) and the use of other adverbs which have a negative weight on this Factor and are, consequently, a sign of situation-dependency, in that they are usually employed for references outside the text (Biber 1988:110).

Table 20. Linguistic features of Dimension 3 of the LSAC

DIMENSION 3

means for conv 17:13 Monday, April 21, 2008 51

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
rel_obj	327	0.3299694	0.4071685	0	3.500000
rel_subj	327	0.5926606	0.6705972	0	5.000000
rel_pipe	327	0.0730887	0.1603598	0	1.400000
p_and	327	0.9067278	0.5830036	0	3.100000
n_nom	327	10.6296636	5.6920088	0	35.400000
tm_adv	327	8.0489297	2.6957574	0	22.200000
pl_adv	327	13.9813456	4.3705850	5.2000000	32.300000
advs	327	56.9284404	8.6095441	34.6000000	99.0000000

The following extract from the LSAC illustrates the dependency of the situation expressed not only by the adverb *tomorrow*, but also by those items (like pronouns) which have a positive weight on Dimension 1:

Extract 7. Features characterizing Dimension 3, from the LSAC

Speaker1: Oh, **she** wants **me** to save **them** for **tomorrow**.

Speaker2: **That** was very good, very good. I especially like **the ones** without sugar **that** you made for me.

Table 21 demonstrates that movie dialogs display extremely similar mean scores regarding the linguistic features just mentioned.

Table 21. Linguistic features of Dimension 3 of the AMC

DIMENSION 3

means per 1,000 words for movies combined = 3 together						17:13 Monday, April 21, 2008
The MEANS Procedure						
Variable	N	Mean	Std Dev	Minimum	Maximum	
fffff	3	0.4666667	0.1527525	0.300000	0.600000	
rel_obj	3	0.6666667	0.3055050	0.400000	1.000000	
rel_subj	3	0.1333333	0.0577350	0.100000	0.200000	
rel_pipe	3	0.6333333	0.2081666	0.400000	0.800000	
p_and	3	13.0333333	3.0730007	10.200000	16.300000	
n_nom	3	6.6333333	0.9712535	5.800000	7.700000	
tm_adv	3	13.6333333	1.0785793	12.400000	14.400000	
pl_adv	3	46.3666667	1.7897858	44.400000	47.900000	
advs	3					

This means that also movie conversation is negative in regard to Dimension 3 and thus relies on situation-dependent reference, like face-to-face conversation, as the following extract shows:

Extract 8. Features characterizing Dimension 3, from the AMC

Speaker1: I have a busy day **today**. Drinks then dinner. Don't wait up will you, darling?

Speaker2: I stopped waiting **a long time ago**, George

Speaker1: Oh and erm, that lunch **tomorrow**, cancel **that** too, will you?

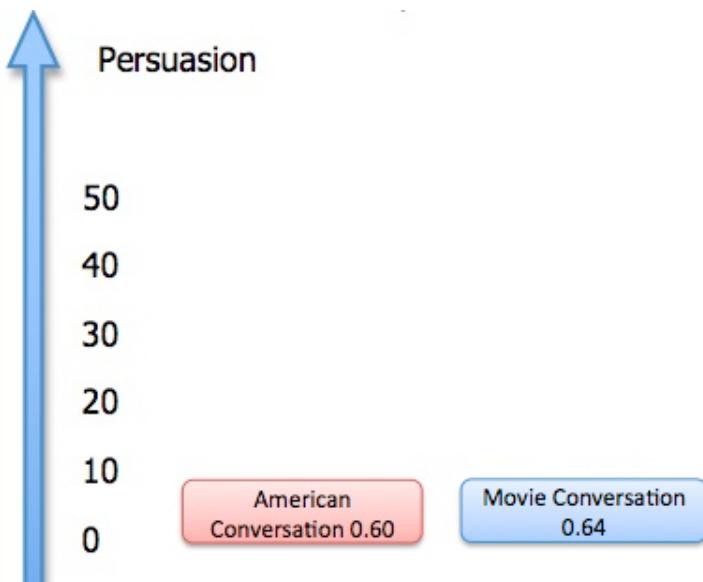
Speaker2: Problems?

Speaker1: I doubt **it**. But Slever Key won't stop calling. You know scientists. They're worse than models. You have to coddle **them** all the time, like little children

4.1.4 Dimension 4: Overt Expression of Persuasion

Also in terms of Dimension 4 (*overt expression of persuasion*, cf. Biber 1988), which has only features with positive weight (cf. Biber, Conrad and Reppen 1998), face-to-face and movie conversation have a very similar positive mean score, namely, 0.64 and 0.60 respectively (cf. Figure 10).

Figure 10. Dimension 4: *overt expression of persuasion*



In linguistic terms, this similarity indicates that both the conversational domains under investigation contain a low percentage of elements that are typical of persuasion: as Table 22 and 23 respectively display, both face-to-face and movie conversation have a low percentage of *infinitive verbs* (*inf* in the table); *modals of prediction* (*will*, *would*, *shall* - *prd_mod*); *suasive verbs* (e.g. *ask*, *command*, *insist* - *sua_vb*); *subordinating conjunctions – conditionals* (e.g. *if*, *unless* - *sub_cnd*); *modals of necessity* (e.g. *ought*, *should*, *must* - *nec_mod*); and *adverbs within auxiliary* (i.e. splitting aux-verb - *spl_aux*) which usually carry weight in persuasive language (Biber 1988). *Infinitive verbs*, for example, can be used as adjectives and verb complements in expressions like *happy to do it*; here, “the head adjective or verb frequently encodes the speaker’s attitude or stance towards the proposition encoded in the infinitival clause” (Biber 1988:111). *Modals* are direct pronouncements that certain events *will* (prediction), *should* (obligation or necessity), *can* or *might* (possibility) occur. *Suasive verbs*, for instance, imply intentions to make an event occur and *conditional subordination* specifies the conditions required to do so. *Split auxiliaries* are often modals, which explains why these features have weight on this dimension (cf. Biber 1988:111).

Table 22. Linguistic features of Dimension 4 of the LSAC

DIMENSION 4

```
means for conv      17:13 Monday, April 21, 2008  51
The MEANS Procedure

Variable      N      Mean      Std Dev      Minimum      Maximum
inf          327    6.8620795    2.3162468      0      19.200000
prd_mod      327    7.0752294    2.9642705      0      19.200000
sua_vb        327    0.2529052    0.6387650      0      9.700000
sub_cnd       327    4.5149847    2.5235556      0      22.200000
nec_mod       327    5.0370031    2.2033454      0      18.900000
spl_aux       327    2.4675841    1.1249763      0      7.300000
```

Table 23. Linguistic features of Dimension 4 of the AMC

DIMENSION 4

```
means per 1,000 words for movies combined = 3 together      17:13 Monday, April 21, 2008
The MEANS Procedure

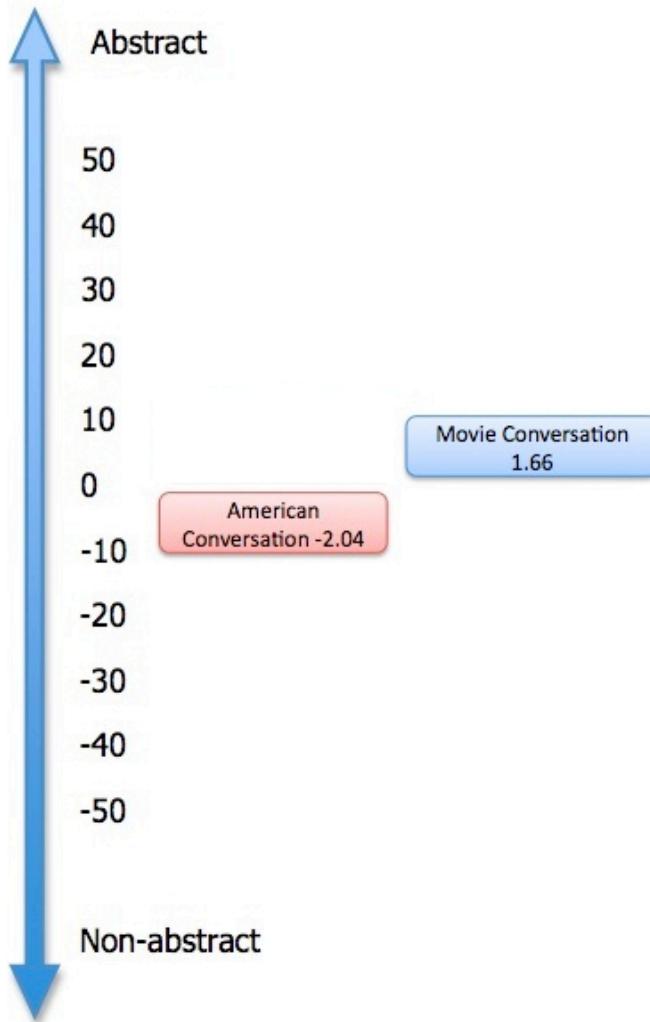
Variable      N      Mean      Std Dev      Minimum      Maximum
inf          3    12.3666667    1.3650397    10.900000    13.600000
prd_mod      3    7.9333333    0.5507571    7.400000    8.500000
sua_vb        3    0.8333333    0.0577350    0.800000    0.900000
sub_cnd       3    3.9000000    0.9539392    3.000000    4.900000
nec_mod       3    5.1333333    0.6429101    4.400000    5.600000
spl_aux       3    2.2333333    0.2081666    2.000000    2.400000
```

The extracts from the corpora above demonstrate that both face-to-face and movie conversation are similar in terms of all the specific features just illustrated.

4.1.5 Dimension 5: Abstract vs. Non-Abstract Information

The only noticeable difference between face-to-face and movie conversation that emerges from Multi-Dimensional analysis concerns Dimension 5 (*abstract versus non-abstract information*; cf. Biber 1988): movie dialog has a positive mean score (1.66) and is, consequently, labeled as *abstract*, whereas face-to-face conversation has a negative mean score (-2.04) and is, consequently, labeled as *non-abstract*. Despite the polar difference, however, it is worth pointing out that the two conversational domains are still extremely similar because the span difference between them is very slight. Figure 11 clearly illustrates their closeness, one either side of 0.

Figure 11. Dimension 5: *abstract versus non-abstract information*



In terms of linguistic features, this similarity is further proved by the rather low percentage of *agentless passive verbs* (*agls_psv* in the tables), of *passive verbs + by* (*by_pasv*), and of *passive postnominal modifiers* (*whiz_vbn*) found in both conversational domains, as the extracts provided above show. These forms are used to reduce emphasis on the agent, to give prominence to the patient of the verb, which is generally an abstract referent (cf. Biber 1988:112).

Looking at Tables 24 and 25, it emerges that the main difference is caused by the slightly higher presence in movies of *conjuncts* (both adverbial, e.g. *however*, *therefore*, *thus* – *conjnts*), *subordination* (e.g. *as*, *except*, *until* – *sub_other*), and *agentless passive verbs* (*agls_psv*). These three features have positive weight on Dimension 5, which has no features bearing heavy negative weight (cf. Biber 1988).

Table 24. Linguistic features of Dimension 5 of the LSAC

DIMENSION 5

means for conv 17:13 Monday, April 21, 2008 51

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
conjnts	327	1.3749235	0.9769653	0	8.8000000
agls_psv	327	3.0431193	1.5035153	0	12.7000000
by_pasv	327	0.1525994	0.2698289	0	2.2000000
whiz_vbn	327	0.4477064	0.6202013	0	7.5000000
sub_othr	327	7.9715596	2.5663499	0	24.9000000

Table 25. Linguistic features of Dimension 5 of the AMC

DIMENSION 5

The MEANS Procedure					
Variable	N	Mean	Std Dev	Minimum	Maximum
conjnts	3	6.8333333	1.3868429	5.3000000	8.0000000
agls_psv	3	4.6333333	0.7505553	3.9000000	5.4000000
by_psv	3	0.2000000	0.1000000	0.1000000	0.3000000
whiz_vbn	3	0.5333333	0.1527525	0.4000000	0.7000000
sub_othr	3	5.1666667	0.1527525	5.0000000	5.3000000

The extracts from the corpora provided above demonstrate that both face-to-face and movie conversation do not have many occurrences of the variables just illustrated.

4.2 Face-to-Face Conversation and Movie Genre

Multi-Dimensional analysis demonstrates that comedies are slightly more similar than non-comedies to face-to-face conversation, even though both types of movies have four Dimensions out of five in common with the latter⁵⁵. As Table 26 shows, comedies have three

⁵⁵ The present Multi-Dimensional analysis is based on comedies and non-comedies only. This is because, since these are more clear-cut than borderline movies, the main interest is to see whether movie genre influences the resemblance, or difference, between face-to-face and movie conversation. However, it has emerged that borderline movies are:

- less involved (Dimension 1) than comedies and non-comedies due to lower occurrences that first person pronouns and possessives, pronoun 'it', discourse particles, adverbial hedges, qualifier – amplifier adverbs, and wh questions have in them and to the higher occurrence of prepositions;
 - more narrative (but still non-narrative and more similar to face-to-face conversation compared to comedies and non-comedies; Dimension 2) than comedies and non-comedies due to the higher occurrences of past tense and public verbs;

Dimensions which are closer to those of face-to-face conversation compared to non-comedies, whereas non-comedies are closer to two. In detail, comedy conversation is more similar to face-to-face conversation with regard to Dimensions 1, 2 and 4, namely, those related to affective, interactional, and generalized contexts, to non-narrative concerns, and to a not particularly high level of persuasion respectively. Non-comedy conversation, instead, is closer to face-to-face conversation as regards Dimensions 3 and 5, namely, those related to situation-dependent factors and to non-abstract information.

Table 26. Comparative Multi-Dimensional analysis of face-to-face conversation versus AMC comedies and AMC non-comedies

Variable	Face-to-Face Conversation	Comedies	Non-Comedies
Dimension 1	35.04	35.86	36.68
Dimension 2	-0.84	-1.11	-1.15
Dimension 3	-7.04	-5.68	-5.93
Dimension 4	0.6	0.24	1.59
Dimension 5	-2.04	2.2	0.87

This closer similarity between comedies and face-to-face conversation becomes more evident by taking into account only the four Dimensions shared with face-to-face conversation (i.e. Dimensions 1, 2, 3, and 4, cf. bold in Table 26). Indeed, by excluding Dimension 5, which is the Dimension on which face-to-face and movie conversation mostly differ, and which neither comedies nor non-comedies share with face-to-face conversation, comedies are closer to face-to-face conversation with regard to three Dimensions (i.e. Dimensions 1, 2, and 4), whereas non-comedies are closer to it only with respect to Dimension 3.

More specifically, non-comedies have a slightly higher occurrence of those features which have a positive weight on Factor 1: as Table 27 illustrates, non-comedies have a higher

-
- less situation dependent (Dimension 3) than comedies and non-comedies due to the higher occurrence of wh clauses and lower occurrence of time adverbs;
 - less characterized by persuasion (Dimension 4) than comedies and non-comedies due to the lower occurrence of modals of necessity and adverbs within auxiliary;
 - abstract (Dimension 5) and their score is half away between comedies and non-comedies due to the fact that the linguistic features characterizing this Factor (i.e. have a score between comedies and non-comedies).
 - most similar in Dimensions 2 and 4 to face-to-face conversation, whereas their Dimension 3 is the most different.

number of *that-deletions* (*that_del* in the tables); *verbs*, in particular *uninflected presents*, *imperatives* and *third persons* (*pres*) and *be* (*be_state*); *it pronouns* (*it*); *causative subordinating conjunctions* (e.g. *because* – *sub_cos*); *wh-questions* (*wh_ques*); and *modals of possibility* (i.e. *can*, *may*, *might*, *could* – *pos_mod*). Conversely, they have a lower occurrence of nouns and attributive adjectives (respectively *n* and *adj_attr*) which have a negative weight on Factor 1. As a consequence, the respective weights of these linguistic items make Factor 1 of non-comedies slightly higher (36.68), and more involved, than that of comedies and face-to-face conversation (35.86 and 35.04 respectively).

Table 27. Linguistic features of Dimension 1 of AMC comedies and non-comedies

DIMENSION 1 (+)

fname	<i>that_del</i>	<i>contrac</i>	<i>pres</i>	<i>pro2</i>	<i>pro_do</i>	<i>pdem</i>	<i>gen_emph</i>	<i>pro1</i>	<i>it</i>
<i>comedies.txt</i>	8.3	7.5	114.7	54.6	4.2	13.0	11.6	75.2	19.0
<i>noncomedies.txt</i>	8.7	5.0	120.4	52.5	4.1	11.7	8.0	72.3	20.5
fname	<i>be_state</i>	<i>sub_cos</i>	<i>prtcle</i>	<i>pany</i>	<i>gen_hdg</i>	<i>amplifr</i>	<i>wh_ques</i>	<i>pos_mod</i>	<i>o_and</i>
<i>comedies.txt</i>	2.7	1.2	8.7	7.1	1.9	2.7	5.6	7.1	11.3
<i>noncomedies.txt</i>	3.7	1.6	7.7	7.7	1.7	2.3	6.4	10.5	11.3
fname	<i>n</i>	<i>prep</i>	<i>adj_attr</i>	<i>typetokn</i>	<i>wrdlnghth</i>				
<i>comedies.txt</i>	194.3	60.5	18.1	48.3	3.8				
<i>noncomedies.txt</i>	189.3	63.0	15.1	56.5	3.9				

The difference related to Dimension 2 indicates that non-comedies are slightly more non-narrative (-1.15) than comedies and face-to-face conversation (-1.11 and -0.84 respectively). This non-narrative feature depends on the fact that either those linguistic items which have a positive weight are less frequent in non-comedies, or those which have a negative weight on this factor are more frequent. By looking at Table 28, which illustrates the features that have positive weight on Dimension 2 (and consequently make the text type more narrative), however, it emerges that the only features which are less frequent in non-comedies are *public verbs* (e.g. *assert*, *complain*, *say*). Indeed, the higher occurrences of *past tense* and *perfect aspect* (*pasttnse* and *perfects* respectively), which have a positive weight on Dimension 2 and consequently make texts more narrative, should make non-comedies more non-narrative than comedies and face-to-face conversation. The opposite result, i.e. the slightly more non-narrative, character of non-comedies, consequently is to be ascribed not only to the slightly higher (4.2-3.2=1.2) occurrences of *public verbs*, but to the higher occurrence of features

which have a negative weight on this Factor. These items happen to be some of the features which have positive weight on Factor 1: by looking back at Table 27 above, it emerges that non-comedies have a higher frequency of *present tense* and *it pronouns*, which not only have weight on Factor 1, making it more involved, but also on Factor 2, making it more non-narrative.

Table 28. Linguistic features of Dimension 2 of AMC comedies and non-comedies

DIMENSION 2 (-)

fname	pasttense	pro3	perfects	pub_vb
comedies.txt	7.1	11.3	2.4	4.4
noncomedies.txt	10.5	11.3	2.6	3.2

As for Dimension 3, non-comedies are more situation-dependent than comedies; this makes the former more similar to face-to-face conversation. As Table 29 demonstrates, this similarity depends on the fact that non-comedies have fewer occurrences of those linguistic items which have a positive weight on Dimension 3, and more occurrences which have a negative weight on it. More specifically, non-comedies have only two out of five linguistic features (i.e. *relative clauses in subject position* and *phrasal connectors*) which carry a negative weight on Factor 3, whereas comedies have three of them (i.e. *relative clause in object position*, *wh pronouns* that function as a relative clause in object position with prepositional fronting, and *nominalization*). Besides, non-comedies have more linguistic items which carry a negative weight (labels bold in Table 29) of Dimension 3 (i.e. time and place adverbs), whereas comedies have only one (the adverbs).

Table 29. Linguistic features of Dimension 3 of AMC comedies and non-comedies

DIMENSION 3 (-)

fname	rel_obj	rel_subj	rel_pipe	p_and	n_nom	tm_adv	pl_adv	advs
comedies.txt	0.3	0.6	0.1	0.8	10.2	6.4	12.4	47.9
noncomedies.txt	0.5	0.4	0.2	0.7	16.3	7.7	14.4	44.4

Dimension 4 is the Factor which displays a major difference in terms of the present

comparison: face-to-face conversation 0.6, comedies 0.24, and non-comedies 1.59. This is translated into face-to-face conversation and comedies being less characterized by persuasion than non-comedies. As illustrated in Table 30, this clearly depends on the higher frequency of the elements which have a positive weight on Dimension 4: *infinitive verbs*, *modals of prediction*, *subordinating conjunctions – conditional*, and *modals of necessity* are, indeed, more frequent in non-comedies; besides, *susative verbs* and *adverbs within auxiliary* are barely higher in comedies.

Table 30. Linguistic features of Dimension 4 of AMC comedies and non-comedies

DIMENSION 4 (+)

fname	inf	prd_mod	sua_vb	sub_cnd	nec_mod	spl_aux
comedies.txt	10.9	7.4	0.8	3.0	5.4	2.3
noncomedies.txt	13.6	8.5	0.9	4.9	5.6	2.4

Dimension 5 is the only Dimension which displays the main difference between face-to-face and movie conversation: movie conversation has a positive mean score (1.66), whereas face-to-face conversation has a negative one (-2.04). As pointed out in the previous section, however, this is not extremely significant because in spite of the polar difference, the span difference between the two conversational domains is very slight. It emerged above that this dissimilarity is mainly caused by the higher presence in movie *conversation of conjuncts*, *subordination*, and *agentless passive verbs*; this is rather interesting for, as Table 31 shows, the highest occurrence of these linguistic items (especially of *agentless passive verbs* and *subordination*) depends mainly on non-comedies, even though they have a mean score which is closer to face-to-face conversation.

Table 31. Linguistic features of Dimension 5 of AMC comedies and non-comedies

DIMENSION 5 (-)

fname	conjnts	agls_psv	by_pasv	whiz_vbn	sub_othr
comedies.txt	8.0	3.9	0.1	0.5	5.0
noncomedies.txt	5.3	5.4	0.3	0.4	5.2

4.3 Discussion of the Multi-Dimensional Results

The Multi-Dimensional analysis of the present chapter has confirmed the findings usually found in the literature on face-to-face conversation (cf. Chapter 2): the data have given further proof that face-to-face conversation has a positive score for Dimension 1, namely, it is characterized by interpersonal, affective, and interactive features; it has a negative score for Dimension 2, namely, it does not have non-narrative concerns; it has a negative score for Dimension 3, namely, it is situation-dependent; it has a positive score for Dimension 4, even though it is not particularly persuasive (cf. Biber 1988), and it has a negative score for Dimension 5, namely, it displays non-abstract information.

With regard to movie conversation, the present approach has re-examined the domain empirically, and what has emerged is a close similarity between movie language and face-to-face conversation. These results are striking in that they contrast the common view in the literature that movie language is artificial and non-spontaneous (cf. Sinclair 2004b, Taylor 1988, Pavese 2005). Indeed, it has emerged from the Multi-Dimensional analysis that face-to-face and movie conversation have more linguistic similarities than differences. More specifically, it has been demonstrated that they both have a positive score as far as Dimension 1 and 4 are concerned, and a negative score with regard to Dimension 2 and 3; the only minimal difference that has been found concerns Dimension 5.

As for Dimension 1, namely, “Informational versus Involved Production” (Biber 1988:107), both movie and spontaneous conversation present a positive factor (i.e. 35.31 and 35.04 respectively). This means that the two domains have a rather high affective, interactional, and generalized content. Indeed, the data have shown that both face-to-face and movie conversation have a high percentage (i.e. more than 50%) of verbs; second person pronouns and possessives; and first person pronouns and possessives.

In terms of Dimension 2, namely “Narrative versus Non-narrative Concerns”, the data have revealed that both spoken conversation and movie dialogs are negative (-0.97 and -0.84 respectively) and are, consequently, characterized by non-narrative concerns, being marked by immediate time and attributive nominal elaboration. Indeed, a relatively low percentage of past tense verbs, of third person pronoun, of verbs in the perfect aspect, and of public verbs has been found.

With regard to Dimension 3, namely “Explicit versus Situation-Dependent Reference”, the data have proved that face-to-face and movie conversation are both negative (-5.7 and -7.04 respectively), in that they both rely on situation-dependent reference: they both have a low percentage (i.e. below 1%) of *wh pronouns* functioning as a relative clause in object position, of *wh pronouns* functioning as a relative clause in subject position, and *wh pronouns* functioning as a relative clause in object position with prepositional fronting.

With respect to Dimension 4, namely “Overt Expression of Persuasion”, the present data have demonstrated that both the conversational domains are positive (0.64 and 0.60 respectively), even though they do not have a high percentage of infinitive verbs, modals of prediction, suasive verbs, subordinating conjunctions, modals of necessity, and adverbs within auxiliary, which have weight on this factor.

The only difference that has emerged from Multi-Dimensional analysis regards Dimension 5, namely “Abstract versus Non-abstract Information”: movie conversation has turned out to have a positive score (1.66) and has, consequently, been defined as *abstract*, whereas face-to-face conversation has a negative score (-2.04) and has, consequently, been labeled as *non-abstract*. Despite this polar difference, however, it has been pointed out that neither of the two conversational domains has a high score (i.e. 1.66 and -2.04 respectively), which means that the difference between them is fairly minimal: a low percentage of agentless passive verbs, passive verbs + *by*, and passive postnominal modifiers characterizes both face-to-face and movie conversation. As a consequence, Dimension 5 has not been considered relevant in differentiating between the two conversational domains; the main difference has been ascribed to the relatively higher presence of adverbial conjuncts in the movies than in face-to-face conversation, namely 6.83 and 1.37 respectively, which have positive weights on this factor. Table 32 summarizes the Multi-Dimensional results (the Dimensions that the two conversational domains have in common are in bold).

Table 32. Summary of the Multi-Dimensional results

MULTI DIMENSIONAL ANALYSIS		
CORPORA:	AMC	LSAC Corpus
DIMENSION 1	(+) Involved Production	
Characterized by linguistic features which contribute to affective, fragmented, interactional, and generalized context, e.g.:		
<ul style="list-style-type: none"> . <i>verbs</i> (uninflected present, imperative and third person) . <i>second person pronouns and possessives</i> . <i>first person pronouns and possessives</i> . <i>private verbs</i> . <i>it pronouns</i> . <i>discourse particles</i> 		
DIMENSION 2	(-) Non-narrative Concerns	
Characterized by linguistic features which contribute to immediate time and attributive nominal elaboration, e.g.:		
<ul style="list-style-type: none"> . present tense . attributive adjectives 		
DIMENSION 3	(-) Situation-Dependent Reference	
<ul style="list-style-type: none"> . Characterized by linguistic features which are usually employed for references outside the text, e.g.: 		
<ul style="list-style-type: none"> . place adverbs . time adverbs 		
DIMENSION 4	(+) (Low) Overt Expression of Persuasion	
Characterized by linguistic features which usually carry weight in persuasive language, e.g.:		
<ul style="list-style-type: none"> . <i>infinitive verbs</i> . <i>modals of prediction</i> . <i>suaive verbs</i> . <i>subordinating conjunctions – conditionals</i> . <i>modals of necessity</i> . <i>adverbs within auxiliary</i> (splitting aux-verb) 		
DIMENSION 5	(+) Abstract	(-) Non-abstract
Higher presence in movies of linguistic features which characterize abstract information, e.g.:		
<ul style="list-style-type: none"> . <i>conjuncts</i> . <i>subordination</i> . <i>agentless passive verbs</i> 		

As for the comedy vs. non-comedy distinction, the present Multi-Dimensional analysis has shown that although both comedies and non-comedies share four Dimensions out of five with face-to-face conversation, comedies resemble it slightly more. In terms of the Multi-Dimensional analysis, this is because their scores are closer on three Dimensions out of five (i.e. Dimensions 1, 2 and 4), whereas non-comedies have a similar score only on Dimension 3, if Dimension 5, which the two movie genres do not have in common with face-to-face conversation, is not taken into account. Interestingly, it also emerged that the main difference between face-to-face and movie conversation mainly depends on the presence of *agentless passive verbs* and *subordination*; these occur more in non-comedies.

CHAPTER 5. MICRO-ANALYSIS

Having established the similarity of face-to-face and movie conversation through a macro-analysis (i.e. the Multi-Dimensional analysis described in Chapter 4), Chapter 5 presents a micro-analysis by concentrating on the DM *you know*. The interest of this derives from the fact that, as illustrated in Chapter 1, *you know* seems to play an important role in speech (Crystal 1988), since it is very frequent in conversation (Kennedy 1998, Biber *et al.* 1999) and is usually described as being part of the core spoken language (McCarthy 1999, Erman 2001). In particular, Section 5.1 illustrates quantitatively the frequency of *you know* in the LSAC and in the AMC, especially by focusing on its plot distribution and on its occurrence as a two-gram. Section 5.2 subsequently narrows the scope by giving details of the DM *you know*: it investigates the position of DM *you know* in the turn and offers a qualitative overview of the functions it displays. Finally, Section 5.3 discusses the results of the micro-analysis.

5.1 Frequency and Plot Analysis of You Know

The data from the LSAC and the AMC demonstrate that *you know* occurs almost twice as frequently in face-to-face than in movie conversation: 5.3 per thousand words in the former and 2.8 in the latter⁵⁶. The plot analysis⁵⁷ of the general distribution of *you know* retrieved with *Wordsmith Tools 4.0* (cf. Figure 12 and 13) well illustrates this numerical discrepancy; it is, however, worth noting that it also shows that the occurrence of *you know* in the two corpora is rather homogeneous, namely, *you know* occurs in various parts (beginning, middle, end) of both the corpora.

⁵⁶ This corresponds to 12,080 occurrences in the LSAC and 293 in the AMC. Since the corpora are of different sizes (the LSAC and the AMC are made up of 2,272,004 and 104,460 words respectively), the occurrences appearing in this chapter are normalized to 1,000 for comparisons which are not based on frequency order. Conversely, for comparisons based on frequency, the occurrences are kept raw.

⁵⁷ Concord dispersion plots illustrate where the search word occurs in a corpus and shows where mention is made most of the search word in the corpus. The plot shows the following labels: *file* (i.e. the source text file-name); *words* (i.e. the number of words in the source text); *hits* (i.e. the number of occurrences of the search-word); *per 1,000* (i.e. how many occurrences per 1,000 words); *dispersion* (i.e., the plot dispersion value, that is the degree to which a set of values are uniformly spread); and *plot* (i.e. a plot showing where the words appear, where the left edge of the plot represents the beginning and the right edge is the end of the file) (Scott 1998:88).

Figure 12. Plot of *you know* in the LSAC

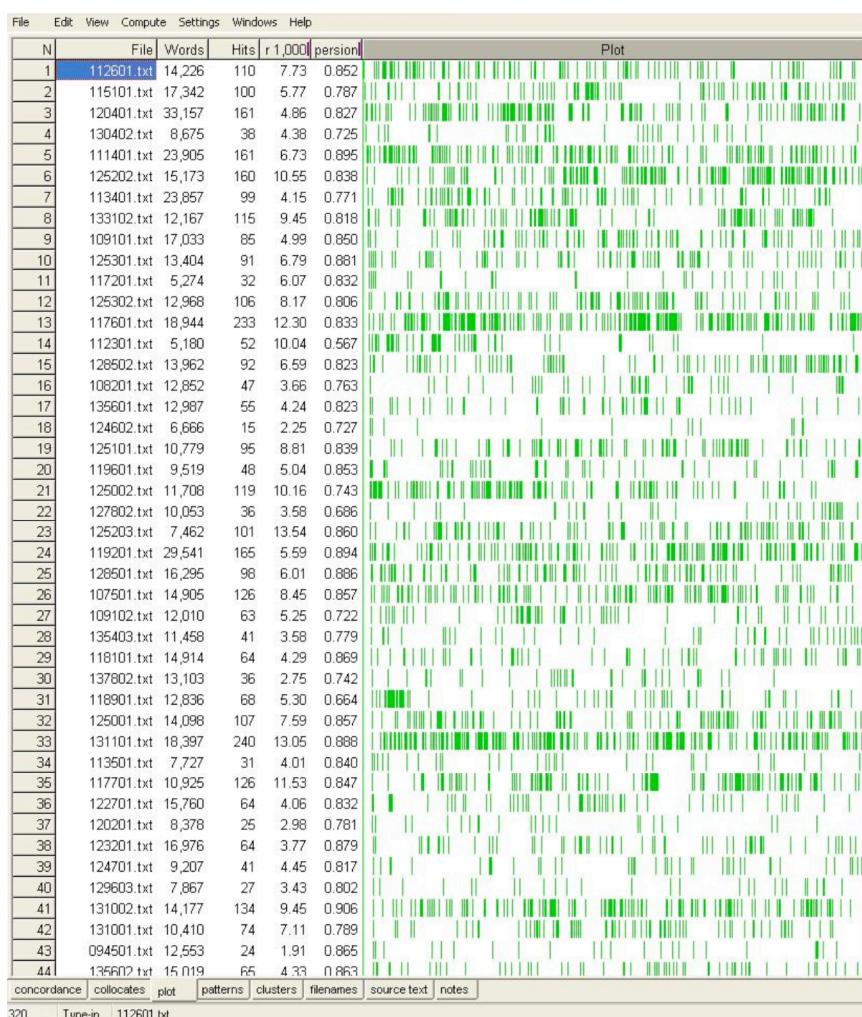
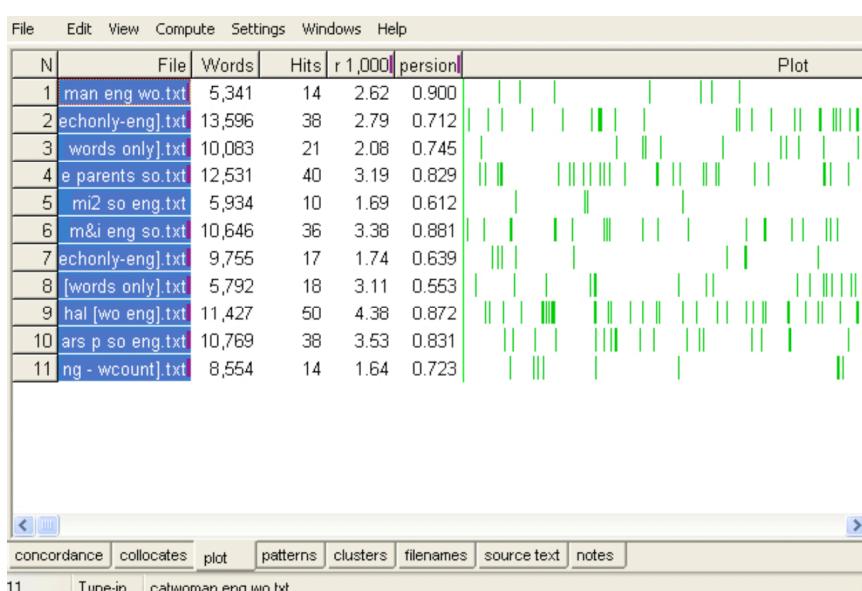


Figure 13. Plot of *you know* in the AMC



It is also important to highlight that despite the numerical discrepancy just illustrated in Figures 12 and 13, the two corpora not only have in common an homogeneous distribution of *you know*, but also an extremely similar patterning concerning its frequency as a two-gram lexical bundle⁵⁸. In Table 33, it clearly emerges that *you know* is the most frequent two-gram present in the two corpora. This is interesting for, as illustrated in Chapter 1, *you know* is one of the core features of spoken language; consequently, its high frequency in both corpora makes movie conversation similar to face-to-face conversation in this regard.

⁵⁸ The focus here is on two-grams only, due to the fact that the DM under investigation, *you know*, is a two-gram.

Table 33. Two-grams present in the LSAC and in the AMC⁵⁹

Face-to-face Conversation			Movie Conversation		
N	Word	Freq.	N	Word	Freq.
1	YOU KNOW	5.32	1	YOU KNOW	2.83
2	I DON'T	3.61	2	I DON'T	2.51
3	IN THE	2.66	3	IN THE	2.44
4	AND I	2.41	4	ARE YOU	2.18
5	I THINK	2.34	5	DO YOU	2.07
6	I MEAN	2.28	6	COME ON	2.01
7	HAVE TO	2.07	7	THIS IS	1.82
8	IT WAS	2.03	8	OF THE	1.81
9	OF THE	2.02	9	ALL RIGHT	1.55
10	AND THEN	2	10	HAVE TO	1.39
11	I WAS	1.98	11	ON THE	1.36
12	GOING TO	1.94	12	I WAS	1.35
13	DON'T KNOW	1.87	13	I HAVE	1.34
14	DO YOU	1.83	14	NO NO	1.31
15	WANT TO	1.57	15	A LITTLE	1.23
16	TO BE	1.54	16	I KNOW	1.22
17	ON THE	1.54	17	THANK YOU	1.22
18	THIS IS	1.47	18	AND I	1.20
19	TO DO	1.45	19	HAVE A	1.14
20	I KNOW	1.43	20	IF YOU	1.14
21	UH HUH	1.36	21	I MEAN	1.12
22	IF YOU	1.31	22	OUT OF	1.12
23	KIND OF	1.31	23	DON'T KNOW	1.11
24	I HAVE	1.29	24	TO DO	1.08
25	YOU HAVE	1.19	25	I THINK	1.06
26	YOU CAN	1.19	26	TO BE	1.04
27	TO THE	1.18	27	I JUST	1.03
28	BUT I	1.16	28	I'M SORRY	1.02
29	HAVE A	1.13	29	TO THE	1.02
30	ARE YOU	1.11	30	YOU HAVE	1

⁵⁹ The numbers in the table are normalized to 1,000.

There are other similarities between the two corpora, concerning the two-grams listed in Table 34: also the second and third most frequent two-grams are identical (cf. *I don't* and *in the* highlighted in azure). Besides, 20 out of the 30 most frequent two-grams in movies are also found within the 30 most frequent two-grams of face-to-face conversation (cf. two-grams highlighted in green); and the other two-grams (e.g. *come on*, *all right*, *no no*, *thank you*) are present in the LSAC, even though they do not occur among its 30 most frequent two-grams. This makes movie conversation closer to face-to-face conversation not only for the linguistic structures they have in common, but also in terms of the pragmatic functions these structures display. Indeed, as illustrated in Chapter 1 and shown in Chapter 4, texts (or, as in this case, corpora) with similar co-occurring linguistic features also share at least one communicative function (Biber 1988:63-64). This claim is further supported by the lexical bundles present in both corpora (i.e. *I don't*, *are you*, *do you*, *come on*, *all right*, *I have*, *thank you*, etc.), which are those which reflect the interpersonal function typical of conversation (cf. Biber 1988, Biber et al. 1991) and highlight the communicative exchange between speakers (Halliday 1993). This is highly relevant for it makes movie conversation pragmatically close to face-to-face conversation.

5.2 *The Discourse Marker You Know*

Due to the high frequency of *you know* in the LSAC (12,080), investigating every occurrence manually to see how many of these occurrences have a discourse marking function was not a realistic option. So, since the issue under examination is a comparison of the functions of *you know* as a discourse marker in two different conversational domains, various strategies had to be adopted to eliminate uses of *you know* which were not those of the discourse marker. It was seen that when *you know* is part of the clusters listed in Table 34, it never has a DM function⁶⁰. Consequently, these clusters were sought in the concordance lines, and the examples of *you know* occurring in these clusters were eliminated. The occurrences left were presumed then to be examples of DM *you know*; indeed, from spot-checks, all the examples turned out to be

⁶⁰ This depends on the fact that in expressions such as *you know what* and *do you know*, for instance, *you know* functions as a full verb: it is used literally to ask the addressee, indirectly (*you know what*) or directly (*do you know*), whether (s)he knows something.

DM *you know*.

Table 34. Non-discourse markers uses of *you know* in the LSAC⁶¹

NON DM <i>you knows</i>	# of occurrences
YOU KNOW WHAT	883
YOU KNOW THAT	358
YOU KNOW HOW	245
YOU KNOW WHEN	135
YOU KNOW SO	128
YOU KNOW WHERE	122
YOU KNOW THAT'S	106
YOU KNOW WHO	82
YOU KNOW WHY	66
YOU KNOW WHAT'S	47
YOU KNOW WHICH	25
TOT 1	2197
DO YOU KNOW	436
THAT YOU KNOW	269
DID YOU KNOW	96
BECAUSE YOU KNOW	94
IF YOU KNOW	52
CAUSE YOU KNOW	51
DON'T YOU KNOW	47
SOMETHING YOU KNOW	37
DIDN'T YOU KNOW	20
TOT 2	1538
TOT 1 + TOT 2	3735

This process resulted in a total of 8,345 occurrences (i.e. 12,080 total occurrences minus 3,735 clusters = 8,345). This number was finally rounded down to 8,000 to eliminate possible miscalculations⁶²; the normalized number of the occurrences of DM *you know* in the

⁶¹ The numbers in the table are not normalized for the aim was to count the actual occurrences so as to subtract them from the total number of *you knows* in the LSAC.

⁶² In the *you know what* counting, for instance, the *do/did you know (what)* cluster might have been included too, thus, summing up the occurrences of both the clusters (to be subtracted from the total number of occurrences of *you know*) could have resulted in counting the same occurrences twice.

whole corpus is 3.52.

As for the AMC, the counting was much easier. This was due to the relatively low occurrences of *you know* (i.e. 162 out of 293 occurrences all checked in context) in the AMC: 1.5 normalized occurrences in the whole corpus.

Another similarity between the two corpora, depicted in Figures 14 and 15, is that the DM uses of *you know* are more frequent than its non-DM uses: 662.25 vs. 552.9 in face-to-face and movie conversation respectively.

Figure 14. Occurrences of the DM and non-DM use of *you know* in the LSAC⁶³

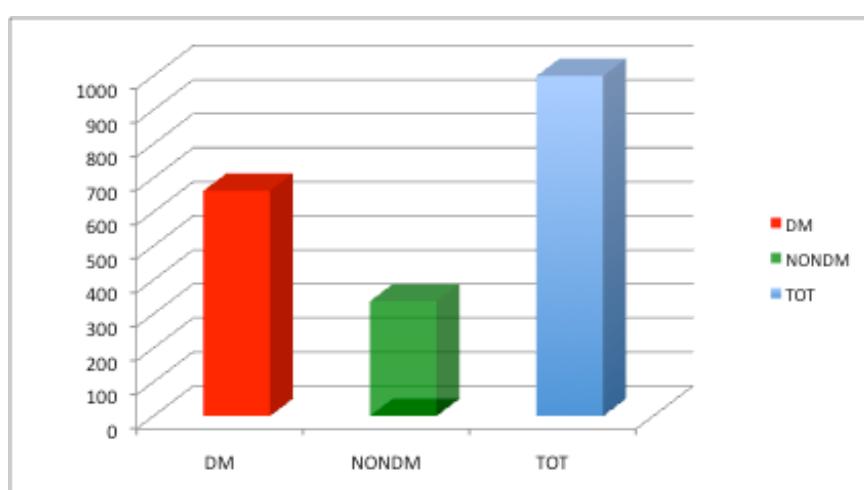
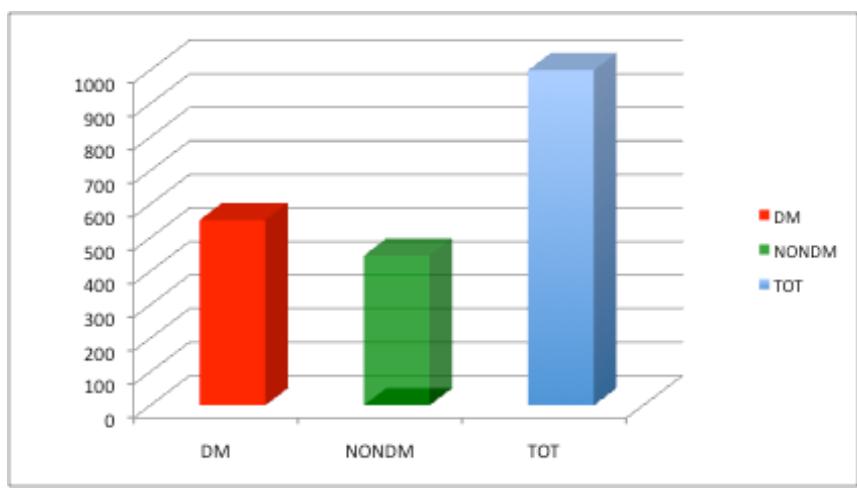


Figure 15. Occurrences of the DM and non-DM use of *you know* in the AMC



⁶³ The numbers in the table are normalized to 1,000.

5.2.1 Turn Position

In order to cope with the large number of occurrences of the DM *you know* in the LSAC (i.e. 8,000), the following analyses of *you know* were based on a sample selection of 165 occurrences randomly chosen from the LSAC (henceforth LSACs). The number of the selection reflects the occurrences of the AMC; the reason for choosing this number does not depend on the fact that it equals the number of occurrences of the DM *you know* in the movies (indeed, norming is usually enough to compare data), but to the fact that nothing new seemed to emerge after checking this sample selection (cf. Chapter 1 on methodology).

In terms of position in the turn, although *you know* can occur in initial, mid and final position, both face-to-face and movie conversation present its highest frequency in mid turn position, fewer occurrences in initial position and rare occurrences in final position (cf. Table 35). This similarity is further backed up by the extremely small numerical difference of the occurrences of *you know* in the two corpora: in mid-position it occurs 715.1 vs. 679 of the time; in initial position 236.3 vs. 265.4 of the time, and in final position 48.4 vs. 55.5 of the time in American and movie conversation respectively.

Table 35. *You know* position in the turn in the LSACs and in the AMC⁶⁴

TURN POSITION	Face-to-face Conversation	Movie Conversation
INITIAL	236.3	265.4
MEDIAL	715.1	679
FINAL	48.4	55.5

5.2.2 Functions

The quantitative analysis based on the data retrieved from the LSAC and the AMC has shown that, even though *you know* occurs with a higher frequency in face-to-face than in movie conversation, it occurs homogeneously, especially in mid-turn position, it is the most frequent two-gram in both the corpora, and its uses as a DM are more frequent than its uses as a non-DM. Moreover, the data have demonstrated that the two conversational domains also

⁶⁴ The numbers in the table are normalized to 1,000.

share similarities concerning other lexical bundles, which serve similar inter-dialogic functions.

In much the same way, also the qualitative analyses based on the investigation of the functions of *you know* in context reveal linguistic similarities: first, they show that the most frequent function of *you know* is the telling one in both corpora; second, that *you know* may occur with other functions, labeled *clarifying*, *time stalling*, and *shared knowledge marking*; third, that even though it may occur with other functions, *you know* displays the telling component also within these other functions; fourth, that *you know* preferably displays the *telling function* in mid position. The following paragraphs give details of these four points that face-to-face and movie conversation have in common.

5.2.2.1 Telling Function

The functions of *you know* were investigated in context, first by checking its occurrences and, then, by analyzing its left and right collocates. As for the former, the occurrences of *you know* in the LSACs show that on nearly three quarters of occasions (715.5), it is used either to provide some (new) information or to comment on something. This is illustrated by examples 1 and 2, where the speaker is providing the listener with new information and a comment respectively (the information/comment is underlined), probably to make sure that the (s)he is following or to make him/her part of the on-going conversation. For this reason, *you know* can be said to occur mostly within a *telling function*.

1. My own I think in those cases you just go on Yeah you know, you just, you can't because sometimes you can't be friends [...] (LSACs)
2. Boy you just came in and were so friendly, you know you just walked in the house like a businessman [...] (LSACs)

In movie dialogs too, more than half the occurrences of *you know* in the AMC have a *telling function* (i.e. 522.2); as examples 3 and 4 demonstrate, *you know* is used to introduce the new information (underlined) provided by the speaker:

3. **You know, Greg's in medicine too.** (AMC)

4. **I need a man around that can give it to me straight, you know?**

Whether the news be good or bad. (AMC)

This frequent *telling function* of *you know* is further confirmed by its right and left collocates in the whole LSAC. This feature emerges from both corpora: Table 36 and 37 indicate that the most frequent collocate one place to the right (R1) of *you know* in both face-to-face and movie conversation is the first person singular, and that other frequent collocates are pronouns and the conjunctions *and* and *but* (cf. bold in the table), referring to people or things (collocates such as *what*, *how*, *when*, *where*, which co-occur with the non-discourse marker *you know* are not taken into account).

Table 36. Fifty most frequent R1 collocates of *you know* in the LSAC

N	Word	With	R1
1	I	YOU KNOW	0.39
2	WHAT	YOU KNOW	0.38
3	AND	YOU KNOW	0.22
4	THE	YOU KNOW	0.16
5	YOU	YOU KNOW	0.15
6	THAT	YOU KNOW	0.15
7	IT'S	YOU KNOW	0.12
8	HOW	YOU KNOW	0.1
9	LIKE	YOU KNOW	0.1
10	IT	YOU KNOW	0.09
11	THEY	YOU KNOW	0.09
12	WE	YOU KNOW	0.09
13	IF	YOU KNOW	0.08
14	HE	YOU KNOW	0.07
15	WHEN	YOU KNOW	0.05
16	SO	YOU KNOW	0.05
17	WHERE	YOU KNOW	0.05
18	BUT	YOU KNOW	0.05
19	I'M	YOU KNOW	0.05
20	THAT'S	YOU KNOW	0.04
21	SHE	YOU KNOW	0.04
22	A	YOU KNOW	0.04
23	UH	YOU KNOW	0.04
24	YEAH	YOU KNOW	0.04
25	THIS	YOU KNOW	0.03
26	WHO	YOU KNOW	0.03
27	JUST	YOU KNOW	0.03
28	IN	YOU KNOW	0.03
29	BECAUSE	YOU KNOW	0.03
30	TO	YOU KNOW	0.03
31	THERE'S	YOU KNOW	0.03
32	HE'S	YOU KNOW	0.03
33	WHY	YOU KNOW	0.02
34	WELL	YOU KNOW	0.02
35	YOU'RE	YOU KNOW	0.02
36	FOR	YOU KNOW	0.02
37	SHE'S	YOU KNOW	0.02
38	THEY'RE	YOU KNOW	0.02
39	SOME	YOU KNOW	0.02
40	WHAT'S	YOU KNOW	0.02
41	ALL	YOU KNOW	0.02
42	CAUSE	YOU KNOW	0.01
43	THOSE	YOU KNOW	0.01
44	THERE	YOU KNOW	0.01
45	MY	YOU KNOW	0.01
46	WE'RE	YOU KNOW	0.01
47	I'VE	YOU KNOW	0.01
48	PEOPLE	YOU KNOW	0.01
49	IS	YOU KNOW	0.01
50	WITH	YOU KNOW	0.01

Table 37. Most frequent R1 collocates of *you know* in the AMC

Word	With	R1
I	YOU KNOW	0.24
THE	YOU KNOW	0.07
YOU	YOU KNOW	0.05
I'M	YOU KNOW	0.04
AND	YOU KNOW	0.03
JUST	YOU KNOW	0.03
HE	YOU KNOW	0.03
IT	YOU KNOW	0.02
WE	YOU KNOW	0.02
IT'S	YOU KNOW	0.02
IF	YOU KNOW	0.02
MY	YOU KNOW	0.02
ONE	YOU KNOW	0.02
BUT	YOU KNOW	0.01
SO	YOU KNOW	0.01
THIS	YOU KNOW	0.01

This high occurrence with the first person singular (and with the other pronouns), indeed, shows that the speaker is providing some information about him/herself (cf. underlined part). This is the case in examples 5 and 6, while examples 7 and 8 refer to other people or things:

5. [...] **you know** I used to go for two, I mean I used to go like for twelve days, you know you span a, a thing and then I, I always like to gave a certain amount. (LSAC)

6. mm I'm **you know** I didn't tape record her, Jack, but that's the impression I got. (AMC)

7. [...] **you know** she drops in. Mm, hmm. So I never, that's why, **you know** it takes me awhile to set her up because I, I never know when she's gone. (LSAC)

8. Oh yeah but **you know** he asked her and she said yes actually he asked my dad. (AMC)

The high occurrence of *you know* with the conjunctions *and* and *but*, instead,

confirms that *you know* is usually employed when a new topic or some new information about a topic is provided, either adding something new (by using *and*) or modifying the previous utterance (by using *but*). These features are illustrated in examples 9 and 12.

9. Uh huh And then Coco was in here **you know and I was telling her**
you know showing her this Right and telling her this and she . (LSAC)

10. I mean, **at first I thought you were in a slum, you know and I could, as a friend,** look the other way while you banged a few fatties and got it out of your system, but **you know** there's lots of good eating fish out there. You don't have to snack on carp any more. (AMC)

11. Speaker1: That thing was tender and ooh, **you know roast is good when it crumbles.**

Speaker2: Yeah. **You know but the roast was that big and it was juicy though. The gravy that came with it was good.** (LSAC)

12. Yeah, that's probably what I should be ordering. **You know but, I don't know, no matter what I eat, my weight just seems to stay the same.** So I figure, what the hell? I'm gonna eat what I want. (AMC)

A similar pattern emerges with regard to the left collocates of *you know*: as listed in Tables 38 and 39, in both the corpora under investigation, the two most frequent L1 collocates of *you know* are *and* and *but* (*do* is not taken into account for it co-occurs with the non-discourse marker *you know*; for the occurrence with other DMs or intlections, instead, see Section 5.2.3).

Table 38. Fifty most frequent L1 collocates of *you know* in the LSAC

N	Word	With	L1
1	AND	YOU KNOW	0.21
2	DO	YOU KNOW	0.19
3	WELL	YOU KNOW	0.17
4	BUT	YOU KNOW	0.15
5	THAT	YOU KNOW	0.11
6	LIKE	YOU KNOW	0.11
7	UH	YOU KNOW	0.09
8	IT	YOU KNOW	0.08
9	SO	YOU KNOW	0.06
10	YEAH	YOU KNOW	0.05
11	UM	YOU KNOW	0.05
12	TO	YOU KNOW	0.04
13	SAID	YOU KNOW	0.04
14	JUST	YOU KNOW	0.04
15	MEAN	YOU KNOW	0.04
16	DID	YOU KNOW	0.04
17	BECAUSE	YOU KNOW	0.04
18	OH	YOU KNOW	0.04
19	THE	YOU KNOW	0.03
20	I	YOU KNOW	0.03
21	YOU	YOU KNOW	0.03
22	IS	YOU KNOW	0.03
23	KNOW	YOU KNOW	0.02
24	THERE	YOU KNOW	0.02
25	OF	YOU KNOW	0.02
26	A	YOU KNOW	0.02
27	WAS	YOU KNOW	0.02
28	THIS	YOU KNOW	0.02
29	OR	YOU KNOW	0.02
30	UP	YOU KNOW	0.02
31	IF	YOU KNOW	0.02
32	THEN	YOU KNOW	0.02
33	CAUSE	YOU KNOW	0.02
34	SAY	YOU KNOW	0.02
35	DON'T	YOU KNOW	0.02
36	ON	YOU KNOW	0.02
37	PEOPLE	YOU KNOW	0.02
38	THING	YOU KNOW	0.02
39	OUT	YOU KNOW	0.01
40	THINK	YOU KNOW	0.01
41	ABOUT	YOU KNOW	0.01
42	STUFF	YOU KNOW	0.01
43	HUH	YOU KNOW	0.01
44	RIGHT	YOU KNOW	0.01
45	SOMETHING	YOU KNOW	0.01
46	SAYING	YOU KNOW	0.01
47	WHAT	YOU KNOW	0.01
48	BE	YOU KNOW	0.01
49	NOT	YOU KNOW	0.01
50	THEM	YOU KNOW	0.01

Table 39. Most frequent L1 collocates of *you know* in the AMC

Word	With	L1
AND	YOU KNOW	0.06
BUT	YOU KNOW	0.05
HAL	YOU KNOW	0.05
HEY	YOU KNOW	0.03
MEAN	YOU KNOW	0.03
WELL	YOU KNOW	0.02
I	YOU KNOW	0.01
UH	YOU KNOW	0.01

Similarly to the right collocates, this high co-occurrence of *and* and *but* to the left of *you know* confirms that the speaker uses *you know* to add some new information to the previous utterance, as in examples 13-16: *and* illustrates that the speaker uses *you know* to add some new information to the previous utterance, whereas *but*, being adversative to the previous statement, provides actual knowledge (i.e. new information) about something by clarifying, re-adjusting or justifying, for example, the previous utterance.

13. And so when they didn't pay their tax they were gonna get in trouble about it, you know, and was going to court and everything and then they were gonna have to pay you know the taxes and so one guy fled the country and he went to Ireland to stay. He left the other one with the bag to hold. So they arrested the other one and the only way he could get out of it and pay his fine and you know, he had to sell his place and sell it real quick. So he sold the house when the property value was eighteen or nineteen thousand dollars. And so our friend here was able to get it real reasonable.
(LSAC)

14. Mmm, delicious thank you. And you know the strawberry's good too.
(AMC)

15. Flying around and he saw in his vision of people in the air fighting but they were fighting with some kind of strange plane and then **you know** he was describing this strange looking thing that was flying around and **but** **you know** we just listened to him, you know. And he was always doing something like that. Now when he finished high school and when he finished college and he was a little school principal by the time he was probably about twenty years old. He <unclear> was a little school principal

out in the country. And like, cause a lot of those girls and boys they were big **you know** and grown up and he was just a little school principal and the only reason he left was to go into the Army... (LSAC)

16. [...] And I could, as a friend, look the other way while you banged a few fatties and got it out of your system, **but you know** there's lots of good eating fish out there. (AMC)

The telling function of *you know* in the cluster *and/but you know* is further proved by the fact that the cluster frequently occurs with *I* and other personal pronouns, which, as shown above, are used to provide new information. The following examples illustrate this:

17. I am so upset because I sat that, I had the children's, the little kids, I was singing them a birthday and thing and I was sending them a little change in it, **you know** not much for the little babies, it was there for his birthday and you know I told Abby today when she called, I said every time I think I'm getting smart and doing something, it makes me so mad it turns out wrong. (LSAC)

18. She stays in one of those projects. But you know they stripped her of her, she can never, I don't know what she did but she can never practice in, in California again. I don't know what she got into but she gets, but you know what she did? (LSAC)

19. Speaker1: Well, I gotta go.

Speaker2: Are you sure? 'Cause my editor for New York Magazine is inside and, you know, I could introduce you two. You sent over your stuff for me to look at? Remember? (AMC)

20. Speaker1: Do you wanna pet the little fella?

Speaker2: No! But you know, I'm not much of a dog person. Uh By the way, you're gonna need a little sod on the fairway there. (AMC)

5.2.2.2 Other Pragmatic Functions

Other common characteristics that face-to-face and movie conversation share emerge from the LSAC and AMC data: first and foremost, the other types of functions which *you know* may occur with; second, the frequency of occurrences of these functions; third, the constant telling component present in all them. Although the presence of the different functions mentioned may suggest a multifunctionality of *you know*, in fact, the functions that *you know* displays are mostly characterized by the fact that when the speaker uses *you know*, some information is always provided.

The other types of functions *you know* may occur with are labeled *clarifying*, *time-stalling* and *shared knowledge marking* function (cf. Chapter 2). In terms of occurrence, there is a difference in the ranking of these functions between the two corpora: in face-to-face conversation, the *clarifying* and the *time-stalling* ones are the second and third most frequent functions (175.7 and 78.7 respectively), whereas in movie conversation the *time-stalling* function is the second most frequent (at 292.9), and the *clarifying* one is in third place (114.6). In both corpora, the least used function is the *shared knowledge marking* one, which occurs in 30.3 of cases in face-to-face conversation, and 67.9 in movie conversation. Figures 16 and 17 illustrate this and also the fact that all the functions constantly occur with the telling one (cf. + *telling* in the picture).

Figure 16. Functions of *you know* in the LSACs

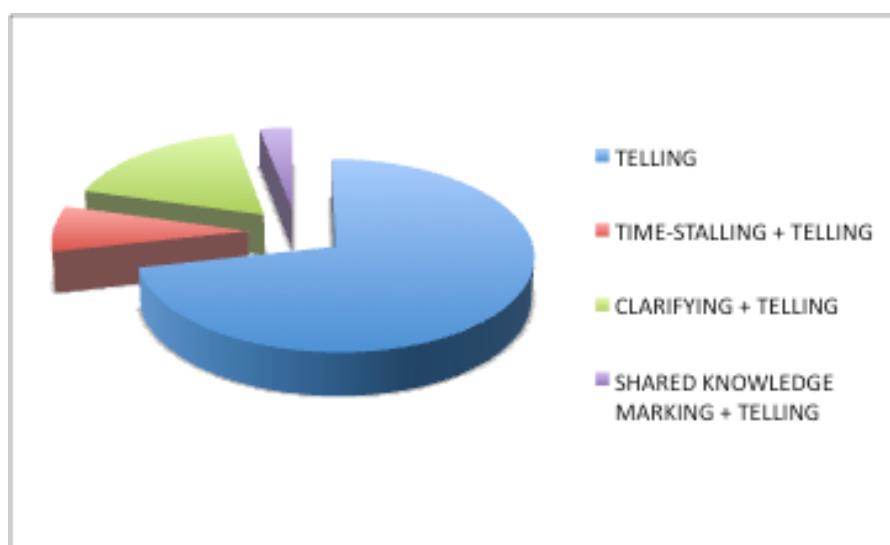
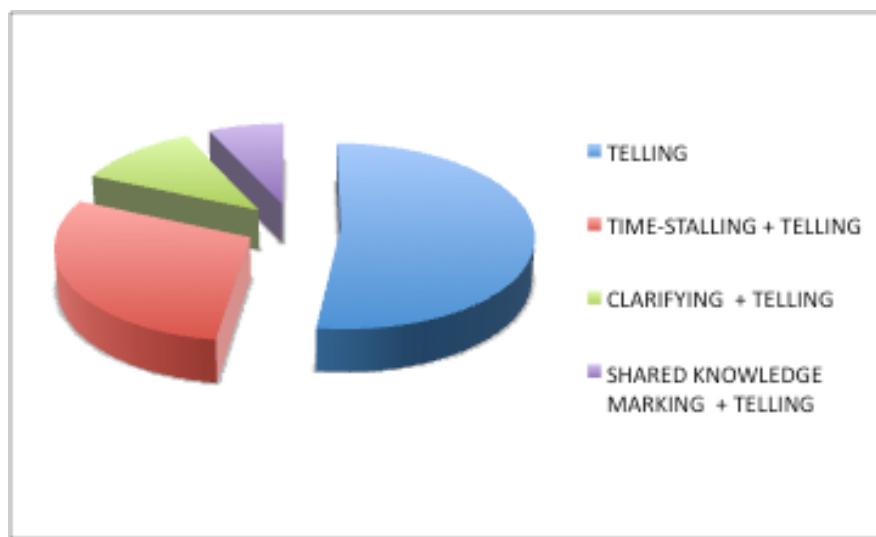


Figure 17. Functions of *you know* in the AMC



The *clarifying function* occurs when the speaker employs *you know* to make his/her statement or a situation more comprehensible; it can be described as occurring with the telling component in that the speaker not only clarifies, but also provides some information about the clarification, as underlined in examples 21-24:

21. Wraps it in like a colorful cellophane, like a you know a a shiny cellophane and puts a big ribbon on it. (LSACs)
22. Playing Colombo with these bozos, you know, that were just shot gunning, you know, a claims process. (LSACs)
23. So, Jack, I mean, couldn't this maybe be construed as illegal? **You know, invasion of privacy or?** (AMC)
24. [...] And in that way, at any time, we can tell the status of a file. You know, where it is in the office, stuff like that. (AMC)

Similarly, the *time-stalling function* occurs when the speaker uses *you know* to try to find the most appropriate expression because (s)he does not know either what to say or how to say it; the telling component is present here for, as underlined in the following examples, the speaker also informs the listener about what (s)he needs and/or thinks, (cf. examples 25 and 26 respectively):

25. I think I just need basil and cumin to make that yeah **you know to make** that uh well the pesto sauce. (LSACs)

26. **You know** I don't believe in this stuff. (AMC)

One feature that emerges in the analysis of *you know* when it is used with a *time-stalling function* is that it occurs with a negative semantic prosody, i.e. in contexts where delicate or unpleasant matters are being talked about, or contexts where the speaker has a negative opinion in some way. This is illustrated in examples 27-30:

27. I'm worried about that guy he's just uh ... to ... you know to many of these heart problems. (LSACs)

28. No, I think they want, **you know** if it's just a **you know** biological problem. (LSACs)

29. **You know**, I think it's a little late for that. (AMC)

30. By not dealing with your problems, Charlie. Ever since, uh Well, **you know**, you've been avoiding confrontation. (AMC)

The delicacy or somehow negative coloring on situations is often emphasized by the co-occurrence of *you know* with other DMs, inserts or hedges used to stall for time and/or slightly soften the situation, as happens in examples 31-38:

31. This sort of thing happens uh, I think uh, with uh Hispanic speakers.
Uh you know, they identify themselves as **uh**, as **you know**. (LSACs)

32. **Well you know** the, it's a problem that can be **uh**, it has mathematics implications and **uh**, that are <unclear> and practical in terms of building machines. <unclear> **I mean** all sorts of words have productive results.
(LSACs)

33. Yes it is. It sure is. **Oh but you know**, know what I'm saying is that...

(LSACs)

34. Yeah. Yeah. **Well you know** it's, it's hard to say exactly when he appeared in America. But it was probably in nineteen twenty-two or twenty-three or something like that though, was it? (LSACs)

35. No, I know that, **but, uhm, yeah, but** this-- this feels different, **you know?** listen jan I can't talk right now. (AMC)

36. **Well**, thank you, Sy, I-- I uh, appreciate that, **but I-- you know**, I really need to get all this stuff and get the hell out of here. (AMC)

37. **You know erm**, I guess there's a few things I need to explain. (AMC)

38. **oh. well, hm.** I'm sorry, Sy. gosh, **you know**, I have to go. hm. it was really nice chatting with you (AMC)

Another sign of negative semantic prosody is suggested by the hesitation⁶⁵ that emerges from both corpora when *you know* co-occurs with repetitions or with syntactic blends, as in shown in examples 39-42:

39. Well **you know the**, it's a problem that can be uh, it has mathematics implications. (LSACs)

40. Well you can eat it **with a, oh I guess** to be out in company, but **you know** you're supposed **to have a, it's a flat fork** (LSACs)

41. I'm dry. **You know, I'm I'm** gonna go get a soda. (AMC)

42. Oh no, no, **I'm just saying is you know come on**, Charlie, goddamn it! (AMC)

It is worth noting that, despite all the signs of an unpleasant situation mentioned so

⁶⁵ Hesitation is being considered here as a sign of negative semantic prosody because of the unpleasant situation it suggests.

far, the *time-staller* *you know* can also be found with positive semantic prosody, especially when the situation is slightly embarrassing because, for example, somebody is paying a compliment (like in example 43) or is inviting somebody out and is afraid of a negative response, as in example 44:

43. **You know** actually I kind of like the table this size. (LSACs)

44. Well, while you're here in town, I mean, **you know**...if you ever feel like taking a break from hanging out with your old sick granny, **you know**, we could... (AMC)

As highlighted above, even though when the *time-staller* *you know* is used, it always implies that there is need for time, it clearly emerges from both corpora that the speaker also provides some information (underlined in the examples) when using it:

45. [She wants] to um do um more studies on um **you know** if, if certain languages are more difficult. (LSACs)

46. Well, uh no. I mean, **you know**, I'd like her to be into culture and shit too. (AMC)

Finally, both in face-to-face and movie conversation, when *you know* occurs within a *shared knowledge marking function*, it implies that the speaker is appealing for or awakening the knowledge (s)he shares with the listener (underlined in the examples). Here again, a telling component emerges within this function. Indeed, *you know* is employed to appeal to knowledge, but, at the same time, also to add some information, as illustrated in examples 47-50:

47. Well **you know** Greg when the kids were little? (LSACs)

48. You want someone that, **you know**, good relation. (LSACs)

49. Speaker1: What do you mean?

Speaker2: **You know, the whole drug thing.** (AMC)

50. Hal, we gotta go to do that thing. **You know, at the at the place.**
(AMC)

The examples so far have demonstrated that although the presence of the different functions mentioned may suggest a multifunctionality of *you know*, the DM is, in fact, constantly characterized by the presence of the telling component. Multifunctionality, however, can be seen to be encountered within the telling component: the speaker in example 51, for instance, is justifying himself by providing information (*telling function*) and clarifying himself (*clarifying function*) (cf. *it's not a normal name* in the example) and, at the same time, he is embarrassed (*time stalling function*) (cf. *I'm sorry, I'm sorry. It's just it's not a* in the example). Similarly, example 52 illustrates that the speaker is providing information, but (s)he needs to find the proper words and clarifies him/herself many times:

51. I'm sorry, I'm sorry. It's just it's not a normal name, **you know.**
(AMC)

52. And this sort of thing happens uh, I think uh, with uh Hispanic speakers. Uh **you know**, they identify themselves as uh, as **you know**, or they identify each other and their friends and their peers and their fellow Spanish speakers by their, by the common language of Spanish. (LSACs)

In much the same way, the speaker in example 53 is embarrassed (cf. *Don't, don't do that. Just, just, uh, let, uh.. you know*) for (s)he has delivered a kind of rude message (i.e. *don't pepper him with questions*); at the same time, he is appealing to knowledge to recover from it (i.e. *you know, people wanna tell their story*):

53. Don't pepper him with questions. Don't, don't do that. Just, just, uh, let, uh.. **you know**, people wanna tell their story. Just let him talk. (AMC)

The speaker in example 53, instead, apart from being embarrassed (i.e. *Uh, I didn't*

wanna tell you before. Well, you know), clarifies the previous part of the message (i.e. *didn't wanna tell you before*) by explaining the reason for not wanting to tell (i.e. *with your worries*):

54. Uh, I didn't wanna tell you before. Well, **you know**, with your worries.
(AMC)

It can be concluded that, despite the multifunctionality usually suggested by the literature, the multifunctionality of *you know* can only be envisaged within its constant telling component. This is especially due to the fact that *you know* usually occurs with a *telling function* and that the other functions *you know* displays are very low in terms of frequency and are mostly used when accompanied by a telling component.

A new term, *co-function*, which recalls Firth's (1957) and Sinclair's (1991) concepts of *collocation* and *colligation* (cf. Chapter 1), is coined here to label this recurrent co-occurrence. *Co-function* is used to mean the likelihood of a lexicogrammatical feature to occur most with a particular function. It can be, consequently, hypothesized that the *telling function* is the *co-function* of *you know*, i.e. the *telling function* is the *functional company* it most frequently keeps, and that the presence of this telling component strongly suggests a basic *core meaning* of the DM analyzed. Indeed, when used with this function, *you know* somehow recalls the semantics of the full verb *to know* which implies knowledge. If this is the case, namely, if the pragmatic meaning of DM is closer to its literal meaning of *you know*, rather than to its non-discourse-marker-like meaning (cf. also Schiffрин 1987), *you know* cannot be said to be grammaticalized, as stated by Chaume (2004b:850)⁶⁶, but rather it is pragmatalized: it does not lose its literal meaning completely, but rather keeps its core meaning within its relative multifunctionality.

5.2.2.3 Functions of *You Know* within its Turn Position

Face-to-face and movie conversation show very similar patterns as regards the functions that *you know* performs according to its position in the turn. As Tables 40 and 41 show, in initial, medial and final position, the most frequent function is the *telling* one (which preferably

⁶⁶ Cf. “*you know* is another discourse marker in the process of grammaticalization” and “its meaning is gradually drifting away from its literal meaning” (Chaume 2004b:850).

occurs in medial position in both corpora and its second most frequent occurrence is in the initial position in both conversational domains).

The most significant difference between face-to-face and movie conversation regards the second most frequent function, which in face-to-face conversation is the *telling + clarifying* one, whereas in movie conversation it is the *telling + time-stalling* one; in both the registers, however, both the functions preferably occurs in mid-position. The least frequent of both the corpora, the *telling + shared knowledge marking* function, occurs in mid-position.

Table 40. Functions of *you know* in the LSACs according to its turn position⁶⁷

FUNCTIONS	INITIAL	MEDIAL	FINAL
TELLING	35	76	7
TELLING + TIME-STALLING	3	10	
TELLING + CLARIFYING		28	1
TELLING + SHARED KNOWLEDGE MARKING	1	4	
TOTAL	39	118	8

Table 41. Functions of *you know* in the AMC according to its turn position

FUNCTIONS	INITIAL	MEDIAL	FINAL
TELLING	22	52	8
TELLING + TIME-STALLING	16	29	1
TELLING + CLARIFYING	2	15	1
TELLING + SHARED KNOWLEDGE MARKING	2	9	
TOTAL	42	105	10

The extremely high frequency of the *telling function* in mid-position may be ascribed to some need to attract and hold the listener's attention while the talk is in progress. It is not at all surprising, though, that *you know* may also acquire a *clarifying function* when it occurs in mid-position; it suggests that the speaker may need to clarify his/her statement while (s)he is talking. In much the same way, the *time-stalling* *you know* in mid-turn position may be used to keep the conversation going by a speaker who cannot find the appropriate words to express his/her ongoing thoughts.

Regarding the functions found in initial position, the most frequent is the *telling* one. This may be because the speaker uses *you know* as a starter to introduce a new topic. The initial *you know*, indeed, may be thought as a kind of rhetorical *you know what/that?* that

⁶⁷ The numbers in the tables are not normalized for the counting is based on the same number of utterances in the two corpora.

implies that the listener does not, in fact, know and needs to be told about something, like in:

55. **You know** you never told me how Aunt Lottie managed to get the drug store. (LSACs)

56. **You know** they were all songs that I thought you like. (LSACs)

57. **You know**, Sean'll never be anything but suspicious if I pitch up saying "Hey honey, I'm home". (AMC)

58. **You know**, I think it's a little late for that. Do give my regards to Gradski, if you see him. (AMC)

Conversely, the final telling *you know* sounds like a *you know that* which anaphorically emphasizes the new information or comment just provided, as illustrated in the following examples:

59. She's gonna blame you, **you know**. (LSACs)

60. Oh I know but I thought you were just saying there wasn't any plugs. No people usually don't have them on the walls. Some people in old houses used to have them cause they would keep their toaster in here **you know**. (LSACs)

61. I just wanted to, um, well, I guess I just wanted to say thank you, **you know?** (AMC)

62. He's right, **you know**. (AMC)

5.2.3 Part of a Bigger Cluster?

Another common characteristic of face-to-face and movie conversation is represented by the common co-occurrence of *you know* with other DMs, interjections, and inserts. In particular, in face-to-face conversation, *you know* especially occurs with *I mean* in the clusters *I mean you*

know and *you know I mean* (respectively 141 and 97 occurrences, see Table 42). This high occurrence suggests that *you know* probably belongs to a larger cluster, which may correspond to the pattern *DM/insert + you know or you know + DM/insert*.

Table 42. Clusters of *you know* in the LSAC

N	Cluster	Freq.
1	YOU KNOW WHAT I	0.11
2	YOU KNOW I MEAN	0.06
3	WHAT I MEAN	0.05
4	YOU KNOW YOU KNOW	0.04
5	DO YOU KNOW WHAT	0.04
6	I MEAN YOU KNOW	0.04
7	YOU KNOW I DON'T	0.04
8	I DON'T KNOW	0.03
9	YOU KNOW AND I	0.03
10	YOU KNOW I THINK	0.03
11	YOU KNOW AND THEN	0.03
12	YOU KNOW I WAS	0.02
13	YOU KNOW IF YOU	0.02
14	YOU KNOW IT'S LIKE	0.02
15	WELL YOU KNOW I	0.02
16	YOU KNOW IT WAS	0.02
17	YOU KNOW WHAT I'M	0.02
18	UH HUH YOU KNOW	0.02
19	YOU KNOW UH HUH	0.02
20	WELL YOU KNOW WHAT	0.02
21	BUT YOU KNOW WHAT	0.02
22	I SAID YOU KNOW	0.02
23	DO YOU KNOW WHERE	0.01
24	AND I SAID	0.01
25	YOU KNOW WHAT YOU	0.01
26	AND YOU KNOW I	0.01
27	AND YOU KNOW WHAT	0.01
28	YOU KNOW I JUST	0.01
29	YOU KNOW I KNOW	0.01
30	BUT YOU KNOW I	0.01

A similar pattern emerges from the movie dialog corpus: even though there are only 0.03 occurrences (i.e. the raw frequency is 4) in the whole AMC of the cluster *you know I mean*, this frequent co-occurrence with other DMs, interjections and inserts is further confirmed by the L1 and R1 collocates of *you know* in both corpora. These collocates are usually expressions like *uh*, *um*, *oh*, *well*, *like*, *yeah*, *just*. In the corpus of face-to-face conversation, there are more R1 collocates than in the AMC, as illustrated in Tables 43 and 44. This may simply be due to a difference in the corpora size or to a difference in genre,

meaning that in movies *you know* might belong only to the *DM/insert + you know* and not to the *you know + DM/insert* cluster, as it does in face-to-face conversation.

Table 43. the 10 most frequent L1 and R1 DM/insert collocates of *you know* in the LSAC

Word	L1	R1
WELL	0.17	
BUT	0.15	0.05
LIKE	0.11	0.10
UH	0.09	0.04
SO	0.06	0.05
YEAH	0.05	0.04
UM	0.05	
JUST	0.04	
MEAN	0.04	0.04
OH	0.04	

Table 44. L1 and R1 DM/insert collocates of *you know* in the AMC

Word	L1	R1
BUT	0.05	0.01
HEY	0.03	
MEAN	0.03	
WELL	0.02	
UH	0.01	
JUST	0.009	0.03
LIKE	0.009	
OH	0.009	

5.2.4 Comedies vs. Non-Comedies: a Matter of Genre?

Pragmatically speaking, in the AMC comedies and non-comedies, *you know* is used in the same way in terms of frequency, functions and functions according to turn position. However, the data from the AMC shows that *you know* occurs twice as frequently in comedies than in non-comedies (there are 104 and 58 occurrences respectively), and the plot analysis retrieved from *Wordsmith Tools 4.0* on the distribution of *you know* in the two genres indicates that this DM has a more homogeneous distribution in comedies (Figure 18) than in non-comedies (Figure 19).

Figure 18. Distribution of *you know* in comedies

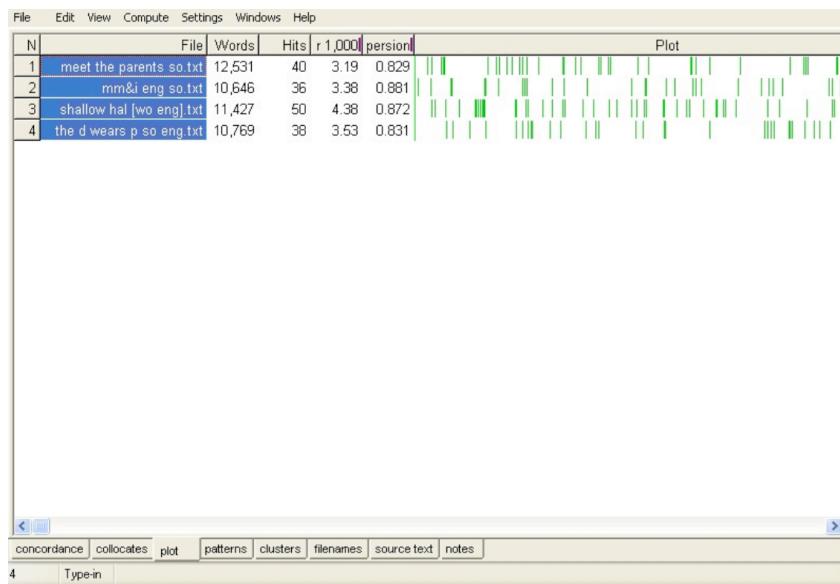
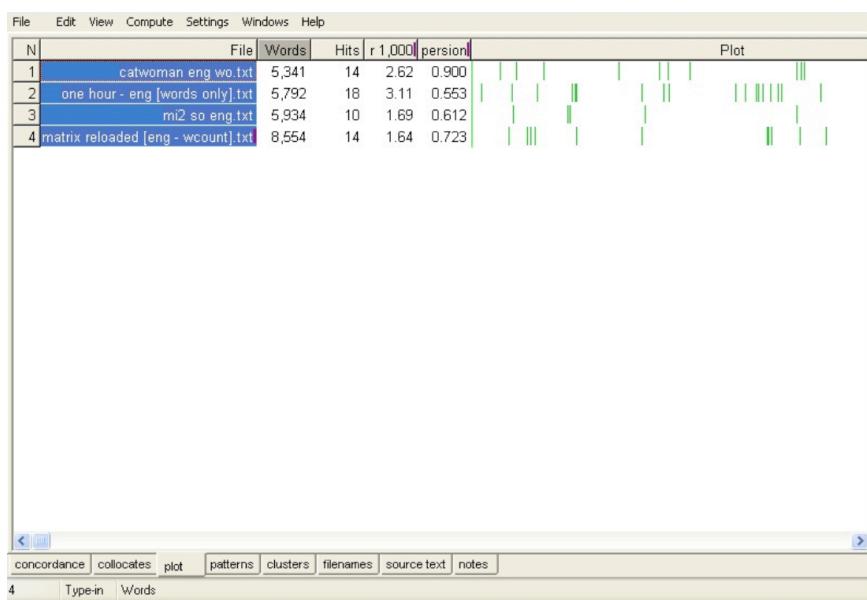


Figure 19. Distribution of *you know* in non-comedies



Naturally, the higher frequency of occurrence of *you know* in comedies has repercussions on the number of its functions; indeed, although the functions that *you know* displays in movies are very similar in the two genres, each function occurs more frequently in comedies than in non-comedies. This is presented in Table 45.

Table 45. Occurrences of the functions of *you know* in comedies and non-comedies⁶⁸

FUNCTIONS	COMEDIES	NONCOMEDIES
TELLING	51	13
TELLING + TIME-STALLING	33	11
TELLING + CLARIFYING	9	3
TELLING + SHARED KNOWLEDGE MARKING	8	2
TOTAL	101	29

In much the same way, also the occurrences of *you know* in the three different turn positions (initial, medial, final) are fewer in non-comedies. What is interesting is that in non-comedies the outstanding occurrence of *you know* in mid-position, typical of face-to-face conversation, is not present. Conversely, its preference of occurrence in mid-position is evident in comedies (cf. Table 46), making them resemble more to face-to-face conversation.

Table 46. Occurrences of *you know* in comedies and non-comedies according to its turn position

POSITION	COMEDIES	NONCOMEDIES
INITIAL	27	12
MEDIAL	67	16
FINAL	7	1
TOTAL	101	29

The same can be said for the functions of *you know* within the turn position: the two genres display similar pragmatic functions, but there are fewer occurrences in non-comedies than in comedies, as Tables 47 and 48 illustrate. The *telling function* in mid-position, especially, which is the most frequent in face-to-face conversation, occurs more in comedies than in non-comedies.

Table 47. Occurrences of the functions of *you know* in comedies according to its turn position

FUNCTIONS	INITIAL	MEDIAL	FINAL
TELLING	13	33	5
TELLING + TIME-STALLING	10	23	
TELLING + CLARIFYING	1	7	1
TELLING + SHARED KNOWLEDGE MARKING	2	6	
TOTAL	26	69	6

⁶⁸ The numbers in the tables of this section are not normalized for the two sub-corpora are already comparable in terms of size.

Table 48. Occurrences of the functions of *you know* in non-comedies according to its turn position

FUNCTIONS	INITIAL	MEDIAL	FINAL
TELLING	7	6	
TELLING + TIME-STALLING	5	5	1
TELLING + CLARIFYING		3	
TELLING + SHARED KNOWLEDGE MARKING		2	
TOTAL	12	16	1

Another feature which comedies share with face-to-face conversation regards the presence of the clusters of *you know*. In particular, the cluster *I mean you know*, discussed above, is present in the AMC comedies, but not in non-comedies (cf. Table 49).

Table 49. *You know* clusters in comedies and non-comedies

MOVIE TYPE	CLUSTERS	#
COMEDIES	YOU KNOW WHAT I'M	7
	YOU KNOW WHAT I	7
	I MEAN YOU KNOW	5
NON-COMEDIES	-	0

The similarity of comedies with face-to-face conversation that emerges from the present data may be ascribed to the fact that conversation might be more important in comedies than in non-comedies, where action counts more. Consequently, when writing comedies, scriptwriters are presumably more careful in planning speech that sounds authentic and spontaneous.

5.3 Discussion of the Micro-Analysis Results

Focusing on movies, the functions found in the AMC have generally confirmed the functions found in the literature (cf. Chapter 3): although *you know* can occur with a *clarifying*, *shared knowledge marking*, and *time-stalling function*, the most frequent function which it usually occurs with is the telling one. The main difference between the present work and the literature regards the *clarifying function*, which in Forchini (*forthcoming*) is said to be the second most frequent function. In the AMC, the clarifying function of *you know* is the third most frequent, while the *time-stalling* function is the second.

Regarding functions in context, the data from the AMC has confirmed that *you know*

occurs with the highest frequency in mid turn position, that it can also occur in initial position, though less frequently, and that it is rare in final position. This has confirmed Erman's (2001) findings, which show that *you know* mostly occurs in mid position and functions as a topic shifter. In the categorization presented here, the topic shifter function is included in the *telling function*.

The data from the present work have confirmed that the DM *you know* occurs both in face-to-face and movie conversation, even though the occurrences present in the former are twice as frequent (3.52) as those present in the latter (1.5). In spite of this numerical difference, the data have proved that the patterning of *you know*, in terms of its general distribution and frequency as a lexical bundle, is extremely similar in the two conversational domains: it occurs homogeneously and especially in turn mid-position; it occurs less in initial position and rarely in final position; and it is the most frequent two-gram present in both the domains empirically investigated. Besides, quantitative and qualitative analyses have also demonstrated that face-to-face and movie conversation share similarities concerning other lexical bundles, which are similar in type and in inter-dialogic function.

The investigation of the functions of *you know* in context has revealed other linguistic similarities between face-to-face and movie conversation: that *you know* preferably and most frequently occurs with the *telling function* in both corpora; that even though *you know* may occur with other functions (i.e. *clarifying*, *time stalling*, and *shared knowledge marking*) too, it displays the telling component also within these other functions; and that *you know* preferably occurs with the *telling function* in mid position.

This preference for the *telling function*, which has been empirically demonstrated by the occurrences of *you know* in context, by both its left and right collocates, and by the telling component co-present in the other functions that *you know* can display, has given proof of a multi-functionality of *you know* which is only apparent, contrary to what is usually suggested by the literature. This, together with the fact that the other functions that *you know* performs occur rarely, has suggested two implications: first and foremost, that *you know* presents a function which is very close to its literal meaning, as already claimed by Schiffrin (1987). Consequently, if this is the case, *you know* cannot be said to be grammaticalized, as stated by Chaume 2004, but rather it can be said to be pragmatalized. It does not completely lose its literal/core meaning, but rather keeps it within its relative multi-functionality. Second, it has emerged that *you know* has a preference of occurrence with a specific function, it has been

claimed that the *telling function* is the *co-function* of *you know*, i.e. the functional company *you know* most frequently keeps. This has been further suggested by the telling component which is co-present in the other functions.

Other similarities which the data have shown regard the functions which *you know* displays according to its position in the turn. More specifically, face-to-face and movie conversation show very similar patterns in that they both display all the functions that *you know* can acquire mostly in mid position; in both conversational domains, the most frequent function in initial, medial and final position is the *telling* one, and the second most frequent occurrences of the *telling function* are in initial position. The most significant difference which has emerged regards the second most frequent function, which in face-to-face conversation is the *clarifying* one, whereas in movie conversation it is the *time-stalling* one; however, in both registers, it preferably occurs in mid-position.

Another shared characteristic between face-to-face and movie conversation is the frequent co-occurrence of *you know* with other DMs, interjections, and inserts. In particular, in face-to-face conversation, *you know* preferably occurs with *I mean* in the clusters *I mean you know* and *you know I mean*, whereas in movie conversation it only occurs in the cluster *I mean you know*. This high occurrence has led to the hypothesis that *you know* may be part of a larger lexical bundle.

The last similarity which has emerged regards the uses of *you know* in different movie genres. Even though the Multi-Dimensional analysis has shown that both comedies and non-comedies share similar linguistic features with face-to-face conversation, and the micro-analysis has confirmed that the functions it displays in the two movie genres are similar, the latter investigation of *you know* has revealed that the former type of movie are closer to spontaneous conversation in terms of frequency (*you know* occurs twice as frequently in comedies than in non-comedies), in terms of the outstanding occurrence of *you know* in mid-position (which is not present in non-comedies), and in terms of the presence of the cluster *I mean you know* (which is absent in non-comedies). This similarity with face-to-face conversation has been ascribed to the fact that in comedies, conversation is probably more important than action, whereas in non-comedies greater weight is placed on action. Consequently, scriptwriters of comedies may pay greater attention to creating speech which sounds more authentic and spontaneous than those of non-comedies, where it is more important that the action holds the audience's attention.

Conclusions

The present dissertation has examined the linguistic features characterizing American face-to-face and movie conversation, two domains which are usually claimed to differ especially in terms of spontaneity. Natural conversation is, indeed, considered the quintessence of the spoken language (Sinclair 2004b) for it is totally spontaneous, whereas movie conversation is usually described as non-spontaneous, being artificially written-to-be spoken (Sinclair 2004b, Taylor 1999, Rossi 2003, Pavesi 2005) and, thus, not likely to represent the general usage of conversation (Sinclair 2004b:80).

The main factor which sparked off the present research was the lack of studies on movie language. Few scholars have actually dealt with transcribed movie material, and have, instead, preferred carrying out a considerable amount of work on movie scripts found on the web (e.g. Taylor 1999, Taylor and Baldry 2004). Others (Baccolini and Bollettieri Bosinelli 1994; Pavesi 1994, 2005; Bollettieri Bosinelli 1998; Pavesi and Malinverno 2000; Taylor 2000a, 2000b, 2000c, 2003; Gottlieb and Gambier 2001; Bruti and Perego 2005; Bruti 2006) have given priority to dubbing and subtitling, rather than to the difference between the two domains mentioned. Besides, some strongly-worded claims about the non-spontaneity of movie language have been based on intuition and not on real movie data (Sinclair 2004b:80). Furthermore, there are no studies of movie language that apply Biber's (1988) Multi-Dimensional analysis, an approach that has proved to be empirically reliable to describe linguistic characteristic of texts.

Taking the factors mentioned into account, the idea of this work was to collect empirical evidence to investigate the extent to which face-to-face and movie conversation actually differ in terms of linguistic features. The main idea was that, given, on the one hand, the increasing insistence on authenticity of spoken data and the complications involved in collecting them (cf. Chapter 1), and, on the other, the relative simplicity of buying movie DVDs and transcribing movie speech, if it could be shown that movie language has similar features to face-to-face conversation, there would be good reasons for using movies for investigating, learning and teaching the spoken language.

For the purpose, the *Longman Spoken American Corpus* (LSAC) was investigated and the *American Movie Corpus* (AMC) was built, explored and finally compared to the former through Multi-Dimensional and micro-analyses.

Following the principles and the methodology illustrated in Chapter 1, after overviewing the general features characterizing face-to-face and movie conversation in Chapter 2 and 3 respectively, Chapters 4 and 5 aimed to answer the following research questions:

- 1) To what extent do face-to-face and movie conversation differ or resemble each other?
- 2) Is *you know*, which has a special status in speech and is part of the core spoken language, equally frequent in movie language?
- 3) What are its pragmatic functions in both face-to-face conversation and movie language, and do these functions vary according to its position in the turn?
- 4) Does movie genre influence this difference or resemblance?

Regarding question number one (to what extent do face-to-face and movie conversation differ or resemble each other?), both the macro and the micro-analysis have empirically shown that the two conversational domains do not differ to a great extent. In particular, the macro Multi-Dimensional analysis demonstrated that they both have a high percentage (i.e. more than 50%) of verbs, second person pronouns/possessives, first person pronouns/possessives, nouns, and prepositions. In much the same way, both face-to-face and movie conversation have an extremely low percentage (i.e. below 1%) of *wh pronouns* (which function as a relative clause), of suasive verbs, passive verbs + *by*, and passive postnominal modifiers, *inter alia*.

In terms of Biber's (1988) Dimensions, it has been empirically shown that both the conversational domains under investigation have a positive score as far as Dimensions 1 and 4 are concerned (which means that they are characterized by high affective, interactional, and generalized content and are not marked overtly by persuasion); negative score with regards to Dimensions 2 and 3 (which means that they are characterized by non-narrative concerns and situation-dependent factors); and that the only difference that emerged (a fairly small one) concerns Dimension 5 (i.e. face-to-face conversation is characterized by abstract information, whereas movie language by non-abstract information). In other words, both face-to-face and movie conversation are informal, non-narrative, situation-dependent and not highly

persuasive. Consequently, since these are all factors linked to the spontaneous nature of conversation (cf. Biber 1988 and Chapter 1), it can then be concluded that also movie language has a significant amount of spontaneity.

As for the micro-analysis and the other research questions (i.e. is *you know* equally frequent in movie language? What are its pragmatic functions in both face-to-face conversation and movie language, and do these functions vary according to its position in the turn?), the data from the LSAC and the AMC have proved that the DM *you know* occurs in both the domains, although the occurrences present in the former are higher (0.35%) than those present in the latter (0.15%). In spite of this numerical difference, the data have proved that face-to-face and movie conversation have an extremely similar patterning of *you know* with respect to its general distribution and to its frequency as a lexical bundle. It occurs homogeneously and especially in turn mid-position; it can also be found, even though less frequently, in initial position, and it is rare in final position. *You know* is also the most frequent 2-gram in both the domains.

Interestingly, the fact that *you know* in both corpora is the most frequent 2-gram and that it occurs homogeneously, in particular in turn mid-position (and less frequently in initial position and rarely in final position), have not been the only traits they have in common. Indeed, in terms of lexical bundles, in particular two-grams, the data have shown that both movie dialogs and spontaneous conversation share other extremely similar features. For instance, the first three most two-grams are identical in the two conversational domains, i.e. *you know, I don't, in the*. Another similar feature is the fact that 20 out of the 30 most frequent two-grams in movies correspond to those present in the 30 most frequent two-grams of face-to-face conversation. Thirdly, other two-grams, such as *come on, all right, no no, thank you*, are present in the American corpus, and reflect the interpersonal character typical of conversation.

The data from both face-to-face and movie conversation have empirically demonstrated that despite its apparent multifunctionality, *you know* preferably occurs with a *telling function*. This preference has been confirmed by the occurrences of *you know* in context, by its left and right collocates, by the telling component present in the other functions, and by the fact that the other functions *you know* may display occur rarely in both corpora. This has further proved the non-grammaticalization of *you know* and the *co-function* (i.e. its preference to occur, a new pragmatic concept introduced here) with the telling

component

As far as functions in context are concerned, the two conversational domains have been shown to be very similar, displaying the *telling function* preferably in mid position. The only difference that has been found is in the second most frequent function of *you know*: the LSAC has a relatively high frequency of the *clarifying function*, whereas the AMC seems to favor the *time-stalling function*.

Another common characteristic which the data have brought to light is the fact that *you know* seems to be part of a larger lexical bundle, for it usually occurs with other DMs, inserts, or interjections. As far as this type of collocates are concerned, only a numerical difference has emerged: in the LSAC there are more R1 collocates than in the AMC. This means that in movies, *you know* may belong to the larger cluster *DM/insert + you know*, whereas in face-to-face conversation, it may belong either to the *DM/insert + you know* or to the *you know + DM/insert* cluster.

Finally, the present work has questioned whether the resemblance encountered between face-to-face and movie conversation may depend on movie genre: even though Multi-Dimensional analysis has shown that both comedies and non-comedies share similar linguistic features with face-to-face conversation, it has emerged that comedies have 3 dimensions out of 5 in common with face-to-face conversation, whereas non-comedies have two. Also, from the investigation of *you know*, comedies have turned out to be closer to spontaneous conversation. Pragmatically speaking, *you know* has been seen to be used in the same way both in comedies and in non-comedies; however, it occurs twice as frequently in comedies as in non-comedies and may be part of a bigger cluster (*I mean you know*) in comedies, but not in non-comedies. This similarity with face-to-face conversation has been ascribed to the fact that in comedies conversation might be more important than action, whereas in non-comedies action counts more. Consequently, scriptwriters presumably plan utterances extremely carefully in comedies, where speech, rather than action, carries a great deal of the movie message.

By way of conclusion, the present research has thus confuted the claim that movie language has “a very limited value” in that it does not reflect natural conversation and, consequently, is “not likely to be representative of the general usage of conversation” (cf. Sinclair 2004b:80). It has thus been shown that movie language can potentially “provide researchers and teachers with a convenient source of spoken language data” (Quaglio and

Biber 2006: 717). However, it must be underlined that the differences between face-to-face and movie conversation, necessarily imposed by the motion-picture medium, cannot be ignored. Indeed, the length of a movie, the necessity to make it appealing to an audience, and create a comprehensible plot, the slower and clearer rhythm of the dialog, the rare occurrences of overlaps and repetitions, and the simplification of structures (all illustrated in Chapter 3) are examples of the non-spontaneity of movies.

The value of this dissertation does not derive only from the fact that it has introduced a new perspective on movie language, namely, its striking similarity with face-to-face conversation, but also from the other novelties that it has brought out. First and foremost, a new corpus was purposely built to study American movie language, the *American Movie Corpus*, without which the present analyses would have not been possible. The AMC is a new type of corpus, because in spite of the relatively large amount of available spoken corpora, there are no corpora of transcribed American movie speech: as mentioned in Chapter 1, web scripts are in fact inappropriate for movie language investigation because a script consistently differs from what is actually said in the movies.

Second, methodologically, the present research has offered a new empirical approach to studying movie language since Multi-Dimensional analysis has never been adopted for research on this type of conversational domain: similar studies based on the Multi-Dimensional approach focused only on television series, not on movies. Examples of these are Rey's (2001), who diachronically investigated male and female language in the American *Star Trek*, and Quaglio's (2004), who compared the language of *Friends* to face-to-face conversation. As a consequence, this new approach, which facilitated investigations that could not have been done manually, has introduced a new perspective on movie language (namely, the great similarity between the two conversational domains), which has provided empirical data, rather than speculations based on intuition.

Third, although the present dissertation has kept the term *discourse marker*, rather than add labels to the already vast list of terms, it has introduced a new description of discourse markers by considering them as *meaningful pragmatic devices, syntactically detachable from the utterance, which connect discourse units by looking forward and/or backward in discourse and which guide the interpretation of the utterance by marking and reinforcing it*. Moreover, it has suggested an approach that uses four umbrella categories to simplify the types of functions with which *you know* may occur. *You know* has been labeled as occurring with a

telling function when it occurs in utterances where the speaker tells or comments on something, often providing new information or information that may be unknown to the listener; with a *clarifying function*, where the speaker makes a statement or a situation more comprehensible either by narrowing and specifying what (s)he means or enlarging and providing further explanation about it; with a *knowledge marking function*, where the speaker appeals for or awakens the knowledge (s)he shares with the listener; and with a *time-stalling function*, where the speaker tries to find the most appropriate expression either because (s)he does not know what to say or how to say it.

Lastly, the concept of a new pragmatic feature has been introduced: the idea of *co-function*, the functional company a lexico-grammatical item most frequently keeps, or the preference of occurrence with a specific function. This has been suggested by the fact that *you know* occurs typically with a very specific function, the telling one, which is present even when *you know* occurs with other functions.

The results of the present research have empirically demonstrated that movie language is extremely close to spontaneous conversation and cannot legitimately be considered of limited value, as Sinclair's (2004b:80) intuition advocated. Although the non-spontaneous traits imposed by the motion-picture medium (cf. Chapter 3), which limit movie language in terms of total spontaneity, cannot be ignored, the main implication of the present results is that the current view of movie dialog as being non-representative of the general usage of conversation needs to be re-considered. The major consequence of this is that movie language can thus be regarded as potentially representative of general conversation usage. In other words, its use as a source of spoken language data for researchers, learners, and teachers becomes legitimate, given that the linguistic features and the level of spontaneity it displays closely resemble face-to-face conversation. Consequently, given the relative ease of collecting movie conversation material, using movies (instead of webscripts) also implies overcoming the complications derived from spoken data collection (cf. Chapter 1).

A further implication which derives from the present research is that thanks to the AMC, it is possible to provide movie language research with empirical data from American movies. Clearly, the AMC is a starting point, since the corpus is still relatively small and needs to be enlarged. However, the advantages of having real movie data have been amply exemplified, and promise that future research on the linguistic features which have emerged from the Multi-Dimensional analyses is needed in further detail.

Undoubtedly, the results of the present dissertation should be verified in other comparable corpora: the first main limitation of this work is, indeed, the fact that corpora cannot capture the whole language system. Besides, the subjectivity of the researcher has to be taken into account: although Multi-Dimensional analysis has been tested in many ways by current linguists, the categorization of *you know* suggested here is subjective; indeed, even though it was based on the analysis of the context in which *you know* locates, data analyses were bound to subjective interpretation. This further emphasizes the need for replicability, which would allow other researchers to check whether the results are consistent or not.

A way to replicate analyses would be to enlarge the AMC, and to check the particular constructions investigated in other corpora. Furthermore, future research could extend the micro-analysis carried out on the discourse marker *you know* (chosen as the most frequent two-gram in both the LSAC and the AMC), to other features of spoken language, to other typical features of spoken conversation. Another aspect which could be developed would be to double-check the hypothesis that *you know* may be part of a larger lexical bundle, to explore the functions it displays in this respect, and to see whether they vary according to the type of DMs, interjections, or inserts it occurs with. Similarly, it would be significant to investigate the new concept of *co-function* introduced here, not only to see whether it is present also in other discourse markers, but especially to see whether it can be applied to other lexical items, as it happens with collocation and colligation; if this were the case, *co-function* would be a further step in the world of the lexico-grammatical interface. It would also be interesting to further investigate movies to see whether other genre categories produce different results. The study could also be broadened out to include and compare other varieties of English. Finally, the AMC or other similar corpora could be used for the aims of language learning, and studies could investigate the effectiveness of using movies with learners to practice their spoken skills, given the fact that it has been demonstrated that movie language can be used to represent conversation.

Appendices

Appendix 1. Linguistic features codes (Multi-Dimensional analysis)

COUNT	Codes	Linguistic Features
{Positive Dimension 1}		
1 =	prv_vb	Private Verbs (e.g. believe, feel, think)
2 =	that_del	'That' Deletion
3 =	contrac	Contraction
4 =	pres	Verb (uninflected present, imperative & third person)
5 =	pro2	Second person pronoun / possessive
6 =	pro_do	Verb 'Do'
7 =	pdem	Demonstrative Pronoun
8 =	gen_emph	Adverb / Qualifier - Emphatic (e.g. just, really, so)
9 =	pro1	First person pronoun / possessive
10 =	it	Pronoun 'it'
11 =	be_state	Verb 'Be' (uninflected present tense, verb and auxiliary)
12 =	sub_cos	Subordinating Conjunction - Causative (e.g. because)
13 =	prtcle	Discourse Particle (e.g. now)
14 =	pany	Nominal Pronoun (e.g. someone, everything)
15 =	gen_hdg	Adverbial - Hedge (e.g. almost, maybe)
16 =	amplifr	Adverb / Qualifier – Amplifier (e.g. absolutely, entirely)
17 =	wh_ques	Wh- question
18 =	pos_mod	Modals of possibility (can, may, might, could)
19 =	o_and	Coordinating conjunction – clausal connector
20 =	wh_cl	Wh- Clause
21 =	finlprep	Stranded Preposition
{Negative Dimension 1}		
22 =	n	Noun
23 =	prep	Preposition
24 =	adj_attr	Attributive Adjective
{Dimension 2}		
25 =	pasttense	Past Tense Verb
26 =	pro3	Third person pronoun (except 'it')

27 =	perfects	Verb – Perfect Aspect
28 =	pub_vb	Public Verbs (e.g. assert, complain, say)
{Dimension 3}		
29 =	rel_obj	Wh pronoun – relative clause – object position
30 =	rel_subj	Wh pronoun – relative clause – subject position
31 =	rel_pipe	Wh pronoun – relative clause – object position with prepositional fronting (' pied piping')
32 =	p_and	Coordinating conjunction – phrasal connector
33 =	n_nom	Singular noun – nominalization
34 =	tm_adv	Adverb – Time
35 =	pl_adv	Adverb – Place
36 =	advs	Adverb (not including counts 8,15,16,34,35,49)
{Dimension 4}		
37 =	inf	Infinitive Verb
38 =	prd_mod	Modal of prediction (will, would, shall)
39 =	sua_vb	Suasive Verb (e.g. ask, command, insist)
40 =	sub_cnd	Subordinating conjunction – conditional (e.g. if, unless)
41 =	nec_mod	Modal of necessity (ought, should, must)
42 =	spl_aux	Adverb within auxiliary (splitting aux-verb)
{Dimension 5}		
43 =	conjnts	Adverbial – conjuncts (e.g. however, therefore, thus)
44 =	agls_psv	Agentless passive verb
45 =	by_pasv	Passive verb + by
46 =	whiz_vbn	Passive postnominal modifier
47 =	sub_othr	Subordinating conjunction – Other (e.g. as, except, until)

Appendix 2. Face-to-face conversation means procedure (Multi-Dimensional analysis)⁶⁹

Variable	N	Mean	Std Dev	Minimum	Maximum
ffffffffffff	327	46.4253823	4.8230528	9.3000000	56.0000000
typetokn	327	3.6941896	0.1260412	3.2000000	4.3000000
wrdlngh	327	5729.45	4299.66	52.0000000	28090.00
wordcnt	327	29.0489297	6.2898343	5.0000000	52.0000000
prv_vb	327	29.4944954	6.2505288	5.6000000	52.2000000
that_del	327	9.8645260	3.3464515	0	23.4000000
contrac	327	2.4464832	1.9780339	0	17.1000000
pres	327	118.2146789	13.8878361	66.5000000	166.7000000
pro2	327	35.3773700	9.3256379	8.4000000	92.4000000
pro_do	327	3.2461774	1.8262677	0	15.0000000
pdem	327	13.1455657	4.2197647	0	29.5000000
gen_emph	327	11.8308869	3.9893793	0	29.3000000
prol	327	65.8051988	13.3580473	25.1000000	100.4000000
it	327	24.6094801	5.8861178	9.2000000	49.7000000
be_state	327	3.2195719	1.6509476	0	11.3000000
sub_cos	327	2.3030581	1.3674760	0	8.0000000
prtcle	327	14.0073394	6.0216894	0	34.8000000
pany	327	7.8688073	2.3328554	0	15.0000000
gen_hdg	327	2.5079511	1.4234876	0	11.0000000
amplifr	327	2.3969419	1.6239739	0	11.1000000
wh_ques	327	3.1651376	1.9488258	0	11.0000000
pos_mod	327	8.3556575	2.8813910	0	24.9000000
o_and	327	8.5972477	3.9154435	0	19.9000000
wh_cl	327	2.5715596	1.5464242	0	19.2000000
finlprep	327	3.3131498	1.6187300	0	14.9000000
n	327	186.4244648	29.7787941	120.6000000	326.2000000
prep	327	63.7510703	10.1261063	0	97.9000000
adj_attr	327	17.5605505	5.2182728	0	42.3000000
pasttnse	327	38.4792049	11.5915373	0	72.2000000
pro3	327	31.9446483	12.7553704	0	96.9000000
perfects	327	5.0917431	2.0774776	0	14.9000000
pub_vb	327	6.5678899	3.1678082	0	23.4000000
rel_obj	327	0.3299694	0.4071685	0	3.5000000
rel_subj	327	0.5926606	0.6705972	0	5.0000000
rel_pipe	327	0.0730887	0.1603598	0	1.4000000
p_and	327	0.9067278	0.5830036	0	3.1000000
n_nom	327	10.6296636	5.6920088	0	35.4000000
tm_adv	327	8.0489297	2.6957574	0	22.2000000
pl_adv	327	13.9813456	4.3705850	5.2000000	32.3000000
advs	327	56.9284404	8.6095441	34.6000000	99.0000000
inf	327	6.8620795	2.3162468	0	19.2000000
prd_mod	327	7.0752294	2.9642705	0	19.2000000
sua_vb	327	0.25229052	0.6387650	0	9.7000000
sub_cnd	327	4.5149847	2.5235556	0	22.2000000
nec_mod	327	5.0370031	2.2033454	0	18.9000000
spl_aux	327	2.4675841	1.1249763	0	7.3000000
conjnts	327	1.3749235	0.9769653	0	8.8000000
agls_psv	327	3.0431193	1.5035153	0	12.7000000
by_pasv	327	0.1525994	0.2698289	0	2.2000000
whiz_vbn	327	0.4477064	0.6202013	0	7.5000000
sub_othr	327	7.9715596	2.5663499	0	24.9000000
vcmp	327	4.2256881	2.4788580	0	38.5000000
downtone	327	1.4911315	0.8965583	0	4.5000000
pred_adj	327	7.6721713	2.9873849	0	38.5000000
allmodal	327	20.4678899	5.2220723	0	43.6000000
allconj	327	24.2993884	6.7531406	6.1000000	46.6000000
allpasv	327	3.6409786	1.6944975	0	12.7000000
allwh	327	5.7376147	2.4953376	0	19.2000000
allwhrel	327	0.9954128	0.8689305	0	6.7000000
alladj	327	40.6474006	8.8416220	15.4000000	76.9000000
allpro	327	133.1229358	17.4763730	74.2000000	172.4000000
have	327	4.6495413	2.0563962	0	16.7000000
allverb	327	148.0642202	11.4394929	104.7000000	178.5000000
vprogrsv	327	12.4703364	3.8513970	0	39.5000000
that_rel	327	2.4785933	1.3357176	0	9.7000000
jcmp	327	0.1229358	0.2163415	0	1.6000000
nonf_vth	327	2.2483180	1.3779114	0	10.2000000
att_vth	327	0.3868502	0.3686048	0	1.9000000
fact_vth	327	4.4311927	2.1048162	0	11.6000000

⁶⁹ Means per 1,000 words.

lkly_vth	327	3.5483180	1.5999967	0	10.8000000
spch_vto	327	0.1678899	0.2741011	0	2.5000000
mntl_vto	327	0.2602446	0.4067799	0	4.4000000
dsre_vto	327	2.4844037	1.2171455	0	7.3000000
efrt_vto	327	1.0385321	0.8348515	0	5.5000000
prob_vto	327	0.1412844	0.2291131	0	1.4000000
x1_jto	327	0.0177370	0.1126524	0	1.6000000
x2_jto	327	0.1125382	0.2849959	0	3.7000000
x3_jto	327	0.0464832	0.1142253	0	0.7000000
x4_jto	327	0.1186544	0.2013468	0	1.4000000
x5_jto	327	0.0972477	0.1619240	0	0.9000000
all_nto	327	0.0366972	0.1164352	0	1.1000000
nonfadvl	327	0.0828746	0.1880159	0	1.9000000
atadvl	327	0.0529052	0.1158423	0	0.7000000
factadvl	327	5.0403670	2.4906495	0	20.1000000
lklyadvl	327	2.3058104	1.7186803	0	22.6000000
all_vth	327	10.6168196	3.3872902	0	22.6000000
all_jth	327	0.1061162	0.1950914	0	1.6000000
all_nth	327	0.1217125	0.2614016	0	3.1000000
all_th	327	10.8434251	3.4247401	0	22.6000000
all_vto	327	4.0886850	1.5333192	0	8.8000000
all_jto	327	0.3923547	0.4037056	0	3.7000000
all_to	327	4.5183486	1.6479490	0	10.5000000
all_advl	327	7.4831804	3.4046781	0	30.1000000
act_ipv	327	0.8431193	0.8261486	0	7.3000000
act_tpv	327	0.8103976	0.6388368	0	4.7000000
mentalpv	327	0.1920489	0.4171159	0	5.3000000
comm_pv	327	0.0088685	0.0532468	0	0.6000000
occur_pv	327	0.0700306	0.1509306	0	1.2000000
copula_pv	327	0.0425076	0.1030091	0	0.9000000
aspect_pv	327	0.2892966	0.4778933	0	4.9000000
humann	327	7.5525994	3.4610964	0	18.4000000
prcessn	327	2.9370031	2.2202652	0	16.7000000
cognitn	327	1.6691131	1.2147332	0	9.9000000
abstrcn	327	9.4923547	3.2952424	0	20.9000000
concrtn	327	10.4498471	4.1987208	0	39.5000000
tccnrcrn	327	1.6544343	1.3290190	0	9.3000000
quann	327	7.5000000	3.0294262	1.2000000	24.9000000
placen	327	4.3149847	2.4430046	0	23.2000000
groupn	327	2.1737003	2.0394678	0	22.6000000
sizej	327	1.9018349	1.2575403	0	9.8000000
timej	327	0.7363914	0.8336426	0	7.4000000
colorj	327	0.3256881	0.5474543	0	5.5000000
evalj	327	1.3226300	0.9094806	0	6.3000000
relatnj	327	1.0954128	0.8697773	0	6.4000000
topicj	327	0.2464832	0.3790263	0	2.4000000
actv	327	37.5076453	7.3615088	14.5000000	74.6000000
commv	327	11.9737003	4.9323067	0	33.0000000
mentalv	327	37.1357798	7.0895740	7.4000000	61.0000000
causev	327	2.1296636	1.4021262	0	9.9000000
occurv	327	1.5308869	1.1545297	0	9.7000000
existv	327	6.5507645	2.2611261	0	17.6000000
aspectv	327	2.0097859	1.1067122	0	8.0000000
all_pv	327	2.2562691	1.2798877	0	10.3000000
epistemic_vth	327	7.9795107	2.7773364	0	17.0000000
concreten	327	12.1042813	4.3149384	0	43.9000000
jandn_th	327	0.2278287	0.3283943	0	3.1000000
abstrctn	327	12.4293578	4.5069266	0	37.6000000
rels	327	3.4740061	1.7965791	0	11.9000000
ffffffffffffffffff					

Appendix 3. Movie conversation means procedure (Multi-Dimensional analysis)⁷⁰

Variable	N	Mean	Std Dev	Minimum	Maximum
ffffffffff	3	53.5333333	4.5456939	48.3000000	56.5000000
typetokn	3	3.8333333	0.0577350	3.8000000	3.9000000
wrdlngth	3	34622.67	9917.29	25556.00	45214.00
wordcnt	3	24.0000000	1.0000000	23.0000000	25.0000000
ttnum	3	24.4000000	0.7211103	23.8000000	25.2000000
prv_vb	3	8.5666667	0.2309401	8.3000000	8.7000000
that_del	3	6.5666667	1.3650397	5.0000000	7.5000000
contrac	3	117.2333333	2.9022979	114.7000000	120.4000000
pres	3	53.3666667	1.0969655	52.5000000	54.6000000
pro2	3	4.3666667	0.3785939	4.1000000	4.8000000
pro_do	3	12.6000000	0.7810250	11.7000000	13.1000000
pdem	3	9.1000000	2.1702534	7.7000000	11.6000000
gen_emph	3	72.3333333	2.8501462	69.5000000	75.2000000
it	3	19.0000000	1.5000000	17.5000000	20.5000000
be_state	3	3.2000000	0.5000000	2.7000000	3.7000000
sub_cos	3	1.4666667	0.2309401	1.2000000	1.6000000
prtcle	3	8.1000000	0.9504385	6.8000000	8.7000000
pany	3	1.5333333	0.4725816	1.0000000	1.9000000
gen_hdg	3	2.3333333	0.3511885	2.0000000	2.7000000
amplifr	3	5.6666667	0.7023769	5.0000000	6.4000000
wh_ques	3	8.4333333	1.8147543	7.1000000	10.5000000
pos_mod	3	11.5666667	0.4618802	11.3000000	12.1000000
o_and	3	2.4666667	0.1154701	2.4000000	2.6000000
wh_cl	3	3.9333333	0.6429101	3.2000000	4.4000000
finlprep	3	191.4666667	2.5658007	189.3000000	194.3000000
n	3	63.4666667	3.2254199	60.5000000	66.9000000
prep	3	16.3000000	1.5874508	15.1000000	18.1000000
adj_attr	3	31.4666667	2.5006666	28.6000000	33.2000000
pasttnse	3	24.2000000	0.6557439	23.5000000	24.8000000
pro3	3	8.8000000	0.3605551	8.4000000	9.1000000
perfects	3	5.2000000	1.1532563	4.3000000	6.5000000
pub_vb	3	0.4666667	0.1527525	0.3000000	0.6000000
rel_obj	3	0.6666667	0.3055050	0.4000000	1.0000000
rel_subj	3	0.1333333	0.0577350	0.1000000	0.2000000
rel_pipe	3	0.6333333	0.2081666	0.4000000	0.8000000
p_and	3	13.0333333	3.0730007	10.2000000	16.3000000
n_nom	3	6.6333333	0.9712535	5.8000000	7.7000000
tm_adv	3	13.6333333	1.0785793	12.4000000	14.4000000
pl_adv	3	46.3666667	1.7897858	44.4000000	47.9000000
advs	3	12.3666667	1.3650397	10.9000000	13.6000000
inf	3	7.9333333	0.5507571	7.4000000	8.5000000
prd_mod	3	0.8333333	0.0577350	0.8000000	0.9000000
sua_vb	3	3.9000000	0.9539392	3.0000000	4.9000000
sub_cnd	3	5.1333333	0.6429101	4.4000000	5.6000000
nec_mod	3	2.2333333	0.2081666	2.0000000	2.4000000
sp1_aux	3	6.8333333	1.3868429	5.3000000	8.0000000
conjcts	3	4.6333333	0.7505553	3.9000000	5.4000000
by_psv	3	0.2000000	0.1000000	0.1000000	0.3000000
whiz_vbn	3	5.1666667	0.1527525	0.4000000	0.7000000
sub_othr	3	1.7000000	0.5196152	1.4000000	2.3000000
vcmp	3	1.4000000	0.4000000	1.0000000	1.8000000
downtone	3	5.2333333	0.7094599	4.6000000	6.0000000
pred_adj	3	21.4666667	2.8005952	19.8000000	24.7000000
allmodal	3	22.8000000	1.2288206	21.4000000	23.7000000
allconj	3	5.3666667	0.8082904	4.5000000	6.1000000
allpasv	3	8.1000000	0.8544004	7.3000000	9.0000000
allwh	3	1.2666667	0.3785939	1.0000000	1.7000000
allwhrel	3	40.1000000	2.9512709	37.2000000	43.1000000
alladj	3	149.9000000	3.2603681	146.8000000	153.3000000
allpro	3	3.9333333	0.1527525	3.8000000	4.1000000
have	3	165.0000000	3.2449961	161.4000000	167.7000000
vprogrsv	3	12.6000000	1.0000000	11.6000000	13.6000000
that_rel	3	0.1666667	0.4932883	1.4000000	2.3000000
jcmp	3	2.2333333	0.4725816	1.7000000	2.6000000
nonf_vth	3	0.6000000	0.2000000	0.4000000	0.8000000
att_vth	3	3.8666667	0.3511885	3.5000000	4.2000000
fact_vth	3	2.6666667	0.4932883	2.1000000	3.0000000

⁷⁰ Means per 1,000 words.

spch_vto	3	0.5000000	0.1000000	0.4000000	0.6000000
mntl_vto	3	0.2666667	0.0577350	0.2000000	0.3000000
dsre_vto	3	3.4666667	0.0577350	3.4000000	3.5000000
efrt_vto	3	1.1333333	0.3214550	0.9000000	1.5000000
prob_vto	3	0.3000000	0	0.3000000	0.3000000
x1_jto	3	0	0	0	0
x2_jto	3	0.1666667	0.1154701	0.1000000	0.3000000
x3_jto	3	0.1000000	0	0.1000000	0.1000000
x4_jto	3	0.1000000	0.1000000	0	0.2000000
x5_jto	3	0.2333333	0.1154701	0.1000000	0.3000000
all_nto	3	0.1000000	0	0.1000000	0.1000000
nonfadvl	3	0.1333333	0.0577350	0.1000000	0.2000000
atadvl	3	0.0666667	0.0577350	0	0.1000000
factadvl	3	4.2333333	0.7234178	3.4000000	4.7000000
lklyadvl	3	1.1000000	0.2645751	0.9000000	1.4000000
all_vth	3	9.3666667	0.6429101	8.9000000	10.1000000
all_jth	3	0.1333333	0.0577350	0.1000000	0.2000000
all_nth	3	0.1333333	0.0577350	0.1000000	0.2000000
all_th	3	9.6000000	0.7000000	9.1000000	10.4000000
all_vto	3	5.6333333	0.4163332	5.3000000	6.1000000
all_jto	3	0.6333333	0.2516611	0.4000000	0.9000000
all_to	3	6.4000000	0.2645751	6.1000000	6.6000000
all_advl	3	5.5000000	0.9643651	4.4000000	6.2000000
act_ipv	3	1.4000000	0.4582576	1.0000000	1.9000000
act_tpv	3	0.7333333	0.2081666	0.5000000	0.9000000
mentalpv	3	0.2333333	0.0577350	0.2000000	0.3000000
comm_pv	3	0	0	0	0
occur_pv	3	0.0333333	0.0577350	0	0.1000000
copulapv	3	0.1000000	0	0.1000000	0.1000000
aspect_pv	3	0.4000000	0.1732051	0.2000000	0.5000000
humann	3	11.6666667	2.0207259	9.5000000	13.5000000
prcessn	3	3.5000000	1.3000000	2.0000000	4.3000000
cognitn	3	2.0666667	0.6658328	1.5000000	2.8000000
abstrcn	3	14.3333333	1.3051181	13.1000000	15.7000000
concrtn	3	10.0666667	1.5534907	8.8000000	11.8000000
tccncrtn	3	2.4333333	0.4041452	2.0000000	2.8000000
quann	3	6.6000000	0.8717798	5.6000000	7.2000000
placen	3	4.0666667	0.9451631	3.0000000	4.8000000
groupn	3	2.3666667	0.3785939	2.1000000	2.8000000
sizej	3	1.2000000	0.6082763	0.8000000	1.9000000
timej	3	0.8333333	0.1527525	0.7000000	1.0000000
colorj	3	0.2000000	0	0.2000000	0.2000000
evalj	3	1.6000000	0.3464102	1.2000000	1.8000000
relatnj	3	0.6666667	0.3511885	0.3000000	1.0000000
topicj	3	0.2000000	0.1000000	0.1000000	0.3000000
actv	3	36.3000000	3.4117444	33.5000000	40.1000000
commv	3	12.6333333	2.2898326	10.8000000	15.2000000
mentalv	3	36.9666667	0.8082904	36.1000000	37.7000000
causev	3	1.8666667	0.3511885	1.5000000	2.2000000
occurv	3	2.2000000	0.6244998	1.7000000	2.9000000
existv	3	7.5333333	1.1015141	6.8000000	8.8000000
aspectv	3	2.1000000	0.3605551	1.8000000	2.5000000
all_pv	3	2.9000000	0.5291503	2.3000000	3.3000000
epistemic_vth	3	6.5333333	0.8144528	5.6000000	7.1000000
concreten	3	12.5000000	1.5588457	11.6000000	14.3000000
jandn_th	3	0.2666667	0.0577350	0.2000000	0.3000000
abstrctn	3	17.8333333	2.4684678	15.1000000	19.9000000
rels	3	3.2333333	0.7637626	2.4000000	3.9000000
ffffffffffffffffff					

Appendix 4. Movie conversation feature counts (Multi-Dimensional analysis)⁷¹

Obs	fname	typetokn	wrdlngh	wordcnt	ttnum	prv_vb	that_del	contrac	pres	pro2
1	bl.txt	55.8	3.8	33098	24	24.2	8.7	7.2	116.6	53.0
2	comedies.txt	48.3	3.8	45214	23	23.8	8.3	7.5	114.7	54.6
3	noncomedies.txt	56.5	3.9	25556	25	25.2	8.7	5.0	120.4	52.5
Obs pro_do pdem gen_emph prol it be_state sub_cos prtcle pany gen_hdg ampliffr wh_ques										
1	4.8	13.1	7.7	69.5 17.5	3.2	1.6	6.8	9.5	1.0	2.0
2	4.2	13.0	11.6	75.2 19.0	2.7	1.2	8.7	7.1	1.9	2.7
3	4.1	11.7	8.0	72.3 20.5	3.7	1.6	7.7	7.7	1.7	2.3
Obs pos_mod o_and wh_cl finlprep n prep adj_attr pasttnse pro3 perfects pub_vb rel_obj										
1	7.7	12.1	2.4	4.2 190.8	66.9	15.7	33.2	24.3	8.9	6.5
2	7.1	11.3	2.4	4.4 194.3	60.5	18.1	32.6	23.5	8.4	4.8
3	10.5	11.3	2.6	3.2 189.3	63.0	15.1	28.6	24.8	9.1	4.3
Obs rel_subj rel_pipe p_and n_nom tm_adv pl_adv advs inf prd_mod sua_vb sub_cnd nec_mod										
1	1.0	0.1	0.4	12.6	5.8	14.1	46.8	12.6	7.9	0.8
2	0.6	0.1	0.8	10.2	6.4	12.4	47.9	10.9	7.4	0.8
3	0.4	0.2	0.7	16.3	7.7	14.4	44.4	13.6	8.5	0.9
Obs spl_aux conjnts agls_psv by_pasv whiz_vbn sub_othr vcmp downtone pred_adj allmodal										
1	2.0	7.2	4.6	0.2	0.7	5.3	1.4	1.4	4.6	19.9
2	2.3	8.0	3.9	0.1	0.5	5.0	1.4	1.0	5.1	19.8
3	2.4	5.3	5.4	0.3	0.4	5.2	2.3	1.8	6.0	24.7
Obs allconj allpasv allwh allwhrel alladj allpro have allverb vprogrsv that_rel jcmp nonf_vth										
1	23.3	5.5	7.3	1.7	37.2	146.8	3.9	167.7	13.6	2.2
2	21.4	4.5	8.0	1.0	43.1	153.3	3.8	161.4	12.6	1.4
3	23.7	6.1	9.0	1.1	40.0	149.6	4.1	165.9	11.6	2.3
Obs att_vth fact_vth lkly_vth spch_vto mntl_vto dsre_vto efrt_vto prob_vto x1_jto x2_jto										
1	0.8	3.5	2.1	0.5	0.3	3.4	1.5	0.3	0	0.1
2	0.4	3.9	3.0	0.4	0.2	3.5	1.0	0.3	0	0.1
3	0.6	4.2	2.9	0.6	0.3	3.5	0.9	0.3	0	0.3
Obs x3_jto x4_jto x5_jto all_nto nonfadvl atadvl factadvl lklyadvl all_vth all_jth all_nth										
1	0.1	0.0	0.1	0.1	0.1	0.1	3.4	0.9	8.9	0.1
2	0.1	0.1	0.3	0.1	0.2	0.1	4.6	1.4	9.1	0.1
3	0.1	0.2	0.3	0.1	0.1	0.0	4.7	1.0	10.1	0.2
Obs all_th all_vto all_jto all_to all_advl act_ipv act_tpv mentalpv commpv occurpv copulapv										
1	9.1	6.1	0.4	6.6	4.4	1.9	0.5	0.3	0	0.0
2	9.3	5.3	0.6	6.1	6.2	1.3	0.9	0.2	0	0.1
3	10.4	5.5	0.9	6.5	5.9	1.0	0.8	0.2	0	0.0
Obs aspectpv humann prcessn cognitn abstrcn concrtn tccncrtn quann placen groupn sizej										
1	0.5	13.5	4.3	1.9	14.2	11.8	2.5	7.0	4.4	2.8
2	0.5	12.0	2.0	1.5	13.1	9.6	2.0	5.6	3.0	2.2
3	0.2	9.5	4.2	2.8	15.7	8.8	2.8	7.2	4.8	2.1

⁷¹ Means per 1,000 words.

Obs	timej	colorj	evalj	relatnj	topicj	actv	commv	mentalv	causev	occurv	existv	aspectv
1	0.7	0.2	1.8	1.0	0.1	40.1	15.2	36.1	1.9	1.7	7.0	2.5
2	1.0	0.2	1.2	0.7	0.2	35.3	11.9	37.7	1.5	2.0	6.8	1.8
3	0.8	0.2	1.8	0.3	0.3	33.5	10.8	37.1	2.2	2.9	8.8	2.0
Obs	all_pv		epistemic_vth	concreten	jandn_th	abstrctn	rels					
1	3.3		5.6		14.3		0.2		18.5		3.9	
2	3.1		6.9		11.6		0.3		15.1		2.4	
3	2.3		7.1		11.6		0.3		19.9		3.4	

References

- Aarts, Bas. 2001. Corpus linguistics, Chomsky and fuzzy tree fragments. In *Corpus linguistics and linguistic theory*, ed. Christian Mair and Marianne Hundt, 5-13. Amsterdam / Atlanta: Rodopi.
- Aarts, Bas and April McMahon. 2006. *The handbook of English linguistics*. Malden / Oxford: Blackwell.
- Aijmer, Karin. 2002. *English discourse particles, evidence from a corpus*. Amsterdam / Philadelphia: John Benjamins.
- Aijmer, Karin and Anna-Brita Stenström. 2005. Approaches to spoken interaction. *Journal of Pragmatics* 37:1743-1751.
- Atkinson, Dwight. 2001. Scientific discourse across history: A combined multi-dimensional/rhetorical analysis of the philosophical transactions of the Royal Society of London. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 45-65. London: Longman.
- Austin, John Langshaw. 1962. *How to do things with words*. Oxford: Clarendon Press.
- Baccolini, Raffaella and Rosa Maria Bollettieri Bosinelli, eds. 1994. *Il doppiaggio: trasposizioni linguistiche e culturali*. Bologna: CLUEB.
- Baker, Mona, Gill Francis, and Elena Tognini-Bonelli, eds. 1993. *Text and technology in honour of John Sinclair*. Philadelphia / Amsterdam: John Benjamins.
- Bazzanella, Carla. 1990. Phatic connectives as intonational cues in contemporary spoken Italian. *Journal of Pragmatics* 14(4):629-647.
- Bazzanella, Carla. 1999. Forme di ripetizione e processi di comprensione nella conversazione. In *La conversazione. Un'introduzione allo studio della conversazione verbale*, ed. Renata Galattolo and Gabriele Pallotti, 205-225. Milano: Raffello Cortina Editore.
- Bercelli, Fabrizio. 1999. Analisi conversazionale e analisi dei frame. In *La conversazione. Un'introduzione allo studio della conversazione verbale*, ed. Renata Galattolo and Gabriele Pallotti, 89-118. Milano: Raffello Cortina Editore.
- Biber, Douglas. 1985. Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics* 23:337-60.
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

- Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistics Computing* 8(4):243-57.
- Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, Douglas. 2004. Conversation text types: A multi-dimensional analysis. *7es Journées internationales d'Analyse statistique des Données Textuelles JADT'04*, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_000.pdf.
- Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam / Philadelphia: John Benjamins.
- Biber, Douglas and Edward Finegan. 1986. An initial typology of English texts. In *New studies in the analysis and exploitation of computer corpora*, ed. Jan Aarts and Eijs Willem, 19-45. Amsterdam: Rodopi.
- Biber, Douglas and Edward Finegan. 2001a. Diachronic relations among speech-based and written registers. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 66-83. London: Longman.
- Biber, Douglas and Edward Finegan. 2001b. Intra-textual variation within medical research articles. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 108-123. London: Longman.
- Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.
- Blakemore, Diane. 1987. *Semantic constraints on relevance*. Oxford: Blackwell.
- Blakemore, Diane. 1992. *Understanding utterances*. Oxford: Blackwell.
- Blakemore, Diane. 2002. *Relevance and linguistic meaning. The semantic and pragmatics of discourse markers*. Cambridge: Cambridge University Press.
- Bolinger, Dwight. 1989. *Intonation and its uses. Melody in grammar and discourse*. London: Edward Arnold.
- Bollettieri Bosinelli, Rosa Maria (ed). 1998. *La traduzione multimediale: quale traduzione per quale testo? Atti del convegno internazionale: La traduzione multimediale*. Bologna: CLUEB.
- Bonomi, Ilaria, Andrea Masini and Silvia Morgana. 2003. *La lingua Italiana e i mass media*.

- Roma: Carocci Editore.
- Börjas, Kersti. 2006. Description and theory. In *The handbook of English linguistics*, ed. Bas Aarts and April McMahon, 9-32. Malden: Blackwell.
- Brinton, Laurel J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.
- Brinton, Laurel J. 2003. I mean: the rise of a pragmatic marker. Paper presented at the Georgetown University Round Table on Languages and Linguistics (GURT), February 15-17, in Georgetown, Washington, D.C.
- Brown, Gillian. 1977. *Listening to spoken English*. London: Longman.
- Brown, Penelope and Stephen Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.
- Bruti, Silvia. 2006. Cross-cultural pragmatics: the translation of implicit compliments in subtitles. *JoSTrans*, Issue 06, http://www.jostrans.org/issue06/art_bruti.php.
- Bruti, Silvia and Elisa Perego. 2005. Translating the expressive function in subtitles: the case of vocatives. In *Research on translation for subtitling in Spain and Italy*, ed. John D. Sanderson, 27-48. Alicante: Publicaciones de la Universidad de Alicante.
- Bubel, Claudia. 2008. Film audience as overhearers. *Journal of Pragmatics* 40:55–71.
- Cameron, Deborah. 2001. *Working with spoken discourse*. London: Sage Publications Ltd.
- Carlson, Lauri. 1984. *'Well' in dialogue games: A discourse analysis of the interjection 'well' in idealized conversation*. Amsterdam / Philadelphia: John Benjamins.
- Carter, Ronald and Michael McCarthy. 2006. *Cambridge grammar of English: A comprehensive guide. Spoken and written English: Grammar and usage*. Cambridge: Cambridge University Press.
- Cattrysse, Patrick. 2001. Multimedia & translation: Methodological considerations. In *(Multi)Media translation. Concepts, practices and research*, ed. Henrik Gottlieb and Yves Gambier, 1-12. Amsterdam / Philadelphia: John Benjamins.
- Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral Literature. In *Spoken and written Language: Exploring orality and literacy*, ed. Deborah Tannen, 35–53. Norwood / New Jersey: Ablex Publishing Corporation.
- Chaume, Frederic. 2004a. *Cine y traducción*. Catedra: Signo e Imagen.
- Chaume, Frederic. 2004b. Discourse markers in audiovisual translating. *Meta* XLIX(4):843–855, <http://www.erudit.org/revue/meta/2004/v49/n4/009785ar.pdf>.

- Chaume, Frederic. 2004c. Film studies and translation studies: Two disciplines at stake in audiovisual translation. *Meta* XLIX(1):12-24,
<http://www.erudit.org/revue/meta/2004/v49/n1/009016ar.pdf>.
- Chomsky, Noam. 1957. *Syntactic structures*. Berlin: Mouton de Gruyter.
- Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: The MIT Press.
- Christoffersen, Erik Exe. 2004. The principle of serendipity. *Nordisk Teaterlaboratorium*,
<http://www.odinteatret.dk/CTL%20web/PDF/WHY%20Christoffersen-EN-pdf.pdf>.
- Clark, Herbert H. and Edward F. Schaefer. 1992. Dealing with overhearers. In *arenas of language use*, ed. Clark Herbert, 248–273. Chicago: University of Chicago Press.
- Conrad, Susan. 2001. Variation among disciplinary texts: a comparison of texts about American nuclear arms policy. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 84-93. London: Longman.
- Contento, Silvana. 1999. Attività Bimodale: aspetti verbali e gestuali della comunicazione. In *La conversazione. Un'introduzione allo studio della conversazione verbale*, ed. Renata Galattolo and Gabriele Pallotti, 267-286. Milano: Raffello Cortina Editore.
- Crystal, David. 1988. Another look at, well, you know... *English Today* 13:47-49.
- Crystal, David and Derek Davy. 1975. *Advanced conversational English*. London: Longman.
- De Saussure, Ferdinand. 1972. *Course in general linguistics*. London: Duckworth.
- Diewald, Gabriele. 2006. Discourse particles and modal particles as grammatical elements. In *Approaches to discourse particles*, ed. Kirsten Fischer, 403-425. Amsterdam: Elsevier.
- Edmondson, Willis. 1981. *Spoken discourse: A model for analysis*. London / New York: Longman.
- Erman, Britt. 1987. *Pragmatic Expressions in English: A study of you know, you see, and I mean in face-to-face conversation*. Stockholm: Almqvist & Wiksell International.
- Erman, Britt. 2001. Pragmatic markers revisited with a focus on ‘you know’ in adult and adolescent talk. *Journal of Pragmatics* 33:1337–1359.
- Firth, John Rupert. 1935a. The technique of semantics. *Transactions of the Philological Society* 36-72.
- Firth, John Rupert. 1935b. The use and distribution of certain English sounds. *English Studies* xvii(I):8-18.
- Firth, John Rupert. 1950. Personality and language in society. *The Sociological Review*

xlii(2):177-189.

- Firth, John Rupert. 1951a. General linguistics and descriptive grammar. *Transactions of the Philological Society* 216-228.
- Firth, John Rupert. 1951b. Modes of meaning. In *Essays and Studies*, ed. John Rupert Firth, 118-149. English Association.
- Firth, John Rupert. 1957a. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, ed. John Rupert Firth et al. 1-32. Special volume of the Philological Society. Oxford: Blackwell.
- Firth, John Rupert. 1957b. *Papers in linguistics 1934-1951*. London: Oxford University Press.
- Fischer, Kerstin. 2006a. Frames, constructions, and morphemic meanings: The functional polysemy of discourse particles. In *Approaches to discourse particles*, ed. Kirsten Fischer, 427-448. Amsterdam: Elsevier.
- Fischer, Kerstin. 2006b. Towards an understanding of the spectrum of approaches to discourse particles: Introduction to the volume. In *Approaches to discourse particles*, ed. Kirsten Fischer, 1-20. Amsterdam: Elsevier.
- Forchini, Pierfranca. *Forthcoming*. Well, uh no. I mean, you know. Discourse markers in movie conversation. In *Lodz Studies in Language*, Frankfurt am Main: Peter Lang.
- Ford, Cecilia A. and Barbara A. Fox, Sandra A. Thompson (ed.). 2002. *The language of turn and sequence*. Oxford: Oxford University Press.
- Fox Tree, Jean E. and Josef C. Schrock. 2002. Basic meaning of 'you know' and 'I mean'. *Journal of Pragmatics* 34:727-47.
- Francis, Gill. 1993. A corpus-driven approach to grammar: principles, methods and examples. In *Text and technology. In honour of John Sinclair*, ed. Mona Baker, Gill Francis and Elena Tognini-Bonelli, 137-156. Amsterdam / Philadelphia: John Benjamins.
- Fraser, Bruce. 1988. Types of English discourse markers. *Acta Linguistica Hungarica* 38:19-33.
- Fraser, Bruce. 1990. An approach to discourse markers. *Journal of Pragmatics* 14:383-95.
- Fraser, Bruce. 1993. Discourse markers across Language. *Pragmatic and Language Learning* (4):1-16. ERIC, ED 396 547.
- Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics* 6:167-190.
- Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics* 31:931-952.
- Fraser, Bruce. 2006. Towards a theory of discourse markers. In *Approaches to discourse particles*, ed. Kirsten Fischer, 189-204. Amsterdam: Elsevier.

- Fuller, Janet. 2003. The influence of speaker roles on discourse marker use. *Journal of Pragmatics* 35:23-45.
- Galattolo, Renata and Gabriele Pallotti, eds. 1999. *La conversazione. Un'introduzione allo studio della conversazione verbale*. Milano: Raffello Cortina Editore.
- Gavioli, Laura. 1999. Alcuni meccanismi di base dell'analisi della conversazione. In *La conversazione. Un'introduzione allo studio della conversazione verbale*, ed. Renata Galattolo and Gabriele Pallotti, 43-66. Milano: Raffello Cortina Editore.
- Goffman, Erving. 1976. Replies and responses. *Language in Society* 5:257–313.
- Goffman, Erving. 1979. Footing. *Semiotica* 25:1–29.
- Gottlieb, Henrik and Yves Gambier, eds. 2001. *Multi-media translation: concepts, practices, and research*. Amsterdam / Philadelphia: John Benjamins.
- Gregory, Michael. 1967. Aspects of varieties differentiation. *Journal of Linguistics* 3:177-98.
- Gregory, Michael and Suzanne Carroll. 1978. *Language and situation: Language varieties and their social contexts*. London: Routledge & Kegan Paul.
- Grice, Paul Herbert. 1975. Logic and Conversation. In *Speech acts* (Syntax and Semantics, Vol. 3), ed. Peter Cole and Jerry L. Morgan, 41-58. New York: Academic Press.
- Halliday, Michael Alexander Kirkwood. 1985a. *An introduction to functional grammar*. London: Arnold.
- Halliday, Michael Alexander Kirkwood. 1985b. *Spoken and written language*. Oxford: Oxford University Press.
- Halliday, Michael Alexander Kirkwood. 1985c. Systemic background. In *Systemic perspectives on discourse, Vol. 1. Selected theoretical papers from the 9th International Systemic Workshop*, ed. James D. Benson and William S. Greaves, (1):1-15. Norwood / New Jersey: Ablex Publishing Corporation.
- Halliday, Michael Alexander Kirkwood. 1992a. Language theory and translation practice. *Rivista Internazionale di Tecnica della Traduzione* (0):15-25.
- Halliday, Michael Alexander Kirkwood. 1992b. Systemic grammar and the concept of a “science of language”. *Waiguoyu (Journal of Foreign Language)* 2(78):1-9.
- Halliday, Michael Alexander Kirkwood. 1993. Quantitative studies and probabilities in grammar. In *Data, description, discourse. Papers on the English language in honour of John Sinclair*, ed. Michael Hoey, 1-25. London: HarperCollins.
- Halliday, Michael Alexander Kirkwood. 1994. *An introduction to functional grammar*.

- London: Arnold.
- Halliday, Michael Alexander Kirkwood. 2003a. Introduction: On the "architecture" of human language. In *On language and linguistics*, ed. Jonathan J. Webster, 1-32. London / New York: Continuum.
- Halliday, Michael Alexander Kirkwood. 2003b (first printed in 1985). Systemic background. In *On language and linguistics*, ed. Jonathan J. Webster, 185-198. London / New York: Continuum.
- Halliday, Michael Alexander Kirkwood. 2003c (first printed in 1992). Systemic grammar and the concept of a "Science of Language". In *On language and linguistics*, ed. Jonathan J. Webster, 199-212. London / New York: Continuum.
- Halliday, Michael Alexander Kirkwood. 2005 (first printed in 2002). The spoken language corpus: a foundation for grammatical theory. In *Computational and quantitative studies*, ed. Jonathan J. Webster, 157-190. London / New York: Continuum.
- Halliday, Michael Alexander Kirkwood and Christian Matthiessen. 2004. *An introduction to functional grammar*. London: Arnold.
- Halliday, Michael Alexander Kirkwood and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
- Hansen, Maj-Britt Mosegaard. 1997. *Alors* and *done* in spoken French: A reanalysis. *Journal of Pragmatics* 28:153-187.
- Hansen, Maj-Britt Mosegaard. 1998. *The function of discourse particles*. Amsterdam: John Benjamins.
- Hawes, Thomas and Sarah Thomas. 1994. Teaching spoken English for informative purposes. *English Teaching Forum*, 32(2):22,
<http://eca.state.gov/forum/vols/vol32/no2/p22.htm>.
- Helt, Marie E. 2001. A multi-dimensional comparison of British and American spoken English. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 171-183. London: Longman.
- Higgins, John. 1991. Looking for patterns. In *Classroom concordancing*, ed. Tim Johns and Philip King, 4:63-70. Birmingham: Birmingham University Press.
- Hoey, Michael. 1991. *Patterns of lexis in text*. Oxford: Oxford University Press.
- Hoey, Michael, ed. 1993. *Data, description, discourse. Papers on the English language in honour of John McH Sinclair*. London: HarperCollins.

- Hoey, Michael. 2005. *Lexical priming. A new theory of words and language*. London / New York: Routledge.
- Hoffmann, Sebastian. 2004. Are low-frequency complex prepositions grammaticalized? On the limits of corpus-data – and the importance of intuition. In *Corpus approaches to grammaticalization in English*, ed. Hans Lindquist and Christian Mair, 171–210. Amsterdam / Philadelphia: John Benjamins.
- Holmes, Janet. 1997. Women, language and identity. *Journal of Sociolinguistics* 1:195–223.
- Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hunston, Susan. 2006. Phraseology and system: A contribution to the debate. In *System and Corpus: Exploring Connections*, ed. Susan Hunston and Geoff Thompson, 55-80. London: Equinox Publishing.
- Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2):249–268.
- Hunston, Susan and Geoff Thompson, eds. 2000. *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.
- Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam / Philadelphia: John Benjamins.
- Izenman, Alan Julian. 2008. *Modern multivariate statistical techniques regression, classification, and manifold learning*. New York: Springer Texts in Statistics.
- James, Allan R. 1983. Compromisers in English: a cross-disciplinary approach to their interpersonal significance. *Journal of Pragmatics* 7:191-206.
- Jefferson, Gail. 1973. A case of precision timing in ordinary conversation. *Semiotica* 9:47–96.
- Johansson, Stig. 1993. Some aspects of the recommendations of the Text Encoding Initiative, with special reference to the encoding of language corpora. In *Corpora Across Centuries*, ed. Merja Kytö, Susan Wright, and Matti Rissanen, 203-210. Amsterdam / Atlanta: Rodopi.
- Johansson, Stig. 2007. Seeing through multilingual corpora. In *Corpus linguistics 25 years on*, ed. Roberta Facchinetto, 51-72. Amsterdam / New York: Rodopi.
- Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London / New York: Longman.
- Labov, William and David Fanshel. 1977. *Therapeutic discourse: Psychotherapy as conversation*. New York: Academic Press.
- Leech, Geoffrey. 1974. *Semantics*. Harmondsworth: Penguin.

- Ler Soon Lay, Vivien. 2006. A Relevance theoretic approach to the discourse particles in Singapore English. In *Approaches to discourse particles*, ed. Kirsten Fischer, 149-166. Amsterdam: Elsevier.
- Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Lewis, Diana. 2006. Discourse markers: A discourse-pragmatic view. In *Approaches to discourse particles*, ed. Kirsten Fischer, 43-60. Amsterdam: Elsevier.
- Lindquist, Hans and Christian Mair, eds. 2004. *Corpus approaches to grammaticalization in English*. Amsterdam / Philadelphia: John Benjamins.
- Lombardo, Linda, Louann Haarman, John Morley, and Christopher Taylor, eds. 1999. *Massed medias. Linguistic tools for interpreting media discourse*. Milano: LED.
- Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and technology. In honour of John Sinclair*, ed. Mona Baker, Gill Francis and Elena Tognini-Bonelli, 157-176. Amsterdam / Philadelphia: John Benjamins.
- Macaulay, Ronald. 2002. You know, it depends. *Journal of Pragmatics* 34:749–67.
- Mahlberg, Michaela. 2006. But it will take time... points of view on a lexical grammar of English. In *The changing faces of corpus linguistics*, ed. Antoinette Renouf and Andrew Kehoe, 377-390. Amsterdam / New York: Rodopi.
- Mair, Christian and Marianne Hundt. 2000. *Corpus linguistics and linguistic theory*. Amsterdam / Atlanta: Rodopi.
- Malinowski, Bronislaw. 1927. The problem of meaning in primitive languages. In *The meaning of meaning*, ed. Charles K. Ogden and Ivor A. Richards (Supplement i):296 – 336. New York: Harcourt, Brace, & Company, inc.
- Malinowski, Bronislaw. 1935. *Coral gardens and their Magic*. London: Allen and Unwin.
- Mansfield, Gillian. 2006. *Changing channels. Media language in (inter)action*. Milano: LED.
- McCarthy, Michael. 1999. What constitutes a basic vocabulary for spoken communication? *Studies in English language and literature* 1:233-249.
- McCarthy, Michael. 2003. *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
- McEnery, Tony and Costas Gabrielatos. 2006. English corpus linguistics. In *The handbook of English linguistics*, ed. Bas Aarts and April McMahon, 33-71. Malden / Oxford: Blackwell.

- McEnery, Tony and Andrew Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.
- Miller, Jim. 2006. Spoken and Written English. In *The handbook of English linguistics*, ed. Bas Aarts and April McMahon, 670-691. Malden / Oxford: Blackwell.
- Miller, Jim and Regina Weinert. 1998. *Spontaneous spoken language*. Oxford: Clarendon.
- Mosegaard Hansen, Maj-Britt. 2006. A Dynamic-Polysemy Approach to the lexical semantics of discourse markers (with an exemplary analysis of French toujours). In *Approaches to Discourse Particles*, ed. Kirsten Fischer, 21-42. Amsterdam: Elsevier.
- Nencioni, Giovanni. 1976. *Di scritto e di parlato. Discorsi linguistici*. Bologna: Zanichelli.
- Norrick, Neal R. 2001. Discourse markers in oral narrative. *Journal of Pragmatics* 33(6):849-878.
- Östman, Jan-Ola. 1981. *You know: A discourse functional approach*. Amsterdam: John Benjamins.
- Partington, Alan. 1998. *Patterns and meanings. Using corpora for English language research*. Amsterdam / Philadelphia: John Benjamins.
- Partington, Alan. 2004. "Utterly content in each other's company": Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1):131-156.
- Pavesi, Maria. 1994. Osservazioni sulla linguistica del doppiaggio. In *Il doppiaggio: trasposizioni linguistiche e culturali*, ed. Raffaella Baccolini and Rosa M. Bollettieri Bosinelli, 129-142. Bologna: CLUEB.
- Pavesi, Maria. 2005. *La Traduzione Filmica. Aspetti del parlato doppiato dall'inglese all'italiano*. Roma: Carocci.
- Pavesi, Maria and Annalisa Malinverno. 2000. Sul turpiloquio nella traduzione filmica. In *Tradurre il cinema*, ed. Christopher Taylor, 75-90. Trieste: La Stea.
- Pons Bordería, Salvador. 2006. A functional approach for the study of discourse markers. In *Approaches to discourse particles*, ed. Kirsten Fischer, 77-100. Amsterdam: Elsevier.
- Quaglio, Paulo. 2004. *The language of NBC's Friends: a comparison with face-to-face conversation*. Unpublished. Ph.D. dissertation, Northern Arizona University
- Quaglio, Paulo and Douglas Biber. 2006. The grammar of conversation. In *The handbook of English linguistics*, ed. Bas Aarts and April McMahon, 692-723. Malden / Oxford: Blackwell.
- Redeker, Gisela. 1991. Linguistic markers of discourse structure. *Linguistics* 29:1139-1172.

- Redeker, Gisela. 2006. Discourse markers as attentional cues at discourse transitions. In *Approaches to discourse particles*, ed. Kirsten Fischer, 339-358. Amsterdam: Elsevier.
- Remael, Aline. 2001. Some thoughts on the study of multimodal, and multimedia translation. In *(Multi)Media Translation. Concepts, practices and research*, ed. Henrik Gottlieb and Yves Gambier, 13-22. Amsterdam / Philadelphia: John Benjamins.
- Renouf, Antoinette. 1997. Teaching corpus linguistics to teachers of English. In *Teaching and language corpora*, ed. Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 255-266. London / New York: Longman.
- Renouf, Antoinette. 2007. Corpus linguistics 25 years on: from super-corpus to cyber-corpus. In *Corpus linguistics 25 years on*, ed. Roberta Facchinetto, 27-50. Amsterdam / New York: Rodopi.
- Renouf, Antoinette and Andrew Kehoe, eds. 2006. *The changing faces of corpus linguistics*. Amsterdam / New York: Rodopi.
- Reppen, Randi. 2001a. Register variation in student and adult speech and writing. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 187-199. London: Longman.
- Reppen, Randi. 2001b. Review of MonoConc Pro and WordSmith Tools. *Language Learning & Technology* 5(3):32-36.
- Reppen, Randi and Nancy Ide. 2004. The American National Corpus: Overall goals and the first release. *Journal of English Linguistics* 32:105-113.
- Rey, Jennifer M. 2001. Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966 to 1993. In *Variation in English: Multi-dimensional studies*, ed. Susan Conrad and Douglas Biber, 138-155. London: Longman.
- Rossi, Alessandra. 2003. La lingua del cinema. In *La lingua italiana e i mass media*, ed. Ilaria Bonomi, Andrea Masini and Silvia Morgana, 93-126. Roma: Carocci Editore.
- Sacks, Harvey. 1992. *Lectures on conversation*. Oxford: Blackwell.
- Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.
- Schiffrin, Deborah. 2001. *Discourse markers: language, meaning, and context*. In *The handbook of discourse analysis*, ed. Deborah Schiffrin, Deborah Tannen, and Heidi Hamilton, 54-75. Oxford: Blackwell.
- Schiffrin, Deborah. 2006. Discourse marker research and theory: Revisiting and. In *Approaches to discourse particles*, ed. Kirsten Fischer, 315-338. Amsterdam: Elsevier.

- Schourup, Lawrence C. 1985. *Common discourse particles in English conversation: like, well, y'know*. New York: Garland.
- Schourup, Lawrence C. 1999. Discourse markers. *Lingua* 107:227-265.
- Scott, Mike. 1998. *WordSmith Tools*. Oxford: Oxford University.
http://www.lexically.net/wordsmith/step_by_step/index.html.
- Scott, Mike and Chris Tribble. 2006. *Textual patterns*. Amsterdam / Philadelphia: John Benjamins.
- Searle, John R. 1969. *Speech acts*. Cambridge: Cambridge University Press.
- Siepmann, Dirk. 2005. *Discourse markers across languages. A contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography*. London / New York: Routledge.
- Sinclair, John McHardy. 1987. *Looking up*. London: Collins.
- Sinclair, John McHardy. 1991. *Corpus concordance collocation*. Oxford: Oxford University Press.
- Sinclair, John McHardy. 1996. The search for units of meaning. *Textus* IX:75–106.
- Sinclair, John McHardy. 1998. The lexical item. In *Contrastive lexical semantics*, ed. Edda Weigand, 1-24. Amsterdam / Philadelphia: John Benjamins.
- Sinclair, John McHardy. 1999. A way with common words. In *Out of corpora: studies in honour of Stig Johansson*, ed. Hilde Hasselgård and Signe Oksefjell, 157-179. Amsterdam: Rodopi.
- Sinclair, John McHardy. 2003. Lexical grammar. *Kompiuterinės Lingvistikos Centras*, <http://donelaitis.vdu.lt/publikacijos/sinclair.pdf>.
- Sinclair, John McHardy. 2004a. *Trust the text: Language, corpus and discourse*. London / New York: Routledge.
- Sinclair, John McHardy. 2004b (first printed in 1987). Corpus creation. In *Corpus linguistics: Readings in a widening discipline*, ed. Geoffrey Sampson and Diana McCarthy, 78-84. London / New York: Continuum.
- Sinclair, John McHardy. 2004c. *Corpus and text: Basic principles on a guide to good practice*. London: AHDS. <http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>.
- Sinclair, John McHardy. 2006. *The case for a Corpus*. Seminar given for the Department of Foreign Languages and Literatures, Università Cattolica del Sacro Cuore, Milan.
- Sperber, Dan and Deidre Wilson. 1986. *Relevance: Communication and cognition*. Oxford:

Blackwell.

- Stame, Stefania. 1999. I marcatori della conversazione. In *La conversazione. Un'introduzione allo studio della conversazione verbale*, ed. Renata Galattolo and Gabriele Pallotti, 169-186. Milano: Raffello Cortina Editore.
- Stern, Karen. 2005. The Longman Spoken American Corpus: providing an in-depth analysis of everyday English, *Pearson Longman*,
<http://www.pearsonlongman.com/dictionaries/pdfs/Spoken-American.pdf>.
- Stubbs, Michael. 1996. *Text and corpus analysis: Computer assisted studies of language and institutions*. Oxford / Massachusetts: Blackwell.
- Stubbs, Michael. 2001. *Words and phrases: Corpus studies in lexical semantics*. Oxford / Massachusetts: Blackwell.
- Stubbs, Michael. 2006. *Quantitative data on multi-word sequences in English: the case of prepositional phrases*. Paper presented at the Berlin-Brandenburgische Akademie der Wissenschaften, 3rd November 2006, in Berlin, Germany.
- Svartvik, Jan. 2007. Corpus linguistics 25 years on. In *Corpus linguistics 25 years on*, ed. Roberta Facchinetti, 11-26. Amsterdam / New York: Rodopi.
- Tannen, Deborah. 1982. The oral/literate continuum in discourse. In *Spoken and written language: Exploring orality and literacy*, ed. Deborah Tannen, 1-16. Norwood / New Jersey: Ablex Publishing Corporation.
- Taylor, Christopher. 1999. Look who's talking. An analysis of film dialogue as a variety of spoken discourse. In *Massed medias. Linguistic tools for interpreting media discourse*, ed. Linda Lombardo, Louann Haarman, John Morley, and Christopher Taylor, 247-278. Milano: LED.
- Taylor, Christopher. 2000a. In Defence of the Word: Subtitles as Conveyors of Meaning and Guardians of Culture. In *La traduzione multimediale. Quale traduzione per quale testo?*, ed. Rosa M. Bollettieri Bosinelli, Christine Heiss, Marcello Soffritti, and Silvia Bernardini, 153-166. Bologna: CLUEB.
- Taylor, Christopher, ed. 2000b. *Tradurre il cinema. Atti del convegno organizzato da G. Soria e C. Taylor 29-30 novembre 1996*. Trieste: Università degli Studi di Trieste.
- Taylor, Christopher. 2000c. The subtitling of film; reaching another community. In *Discourse and community; doing functional linguistics*, ed. Eija Ventola, 309-330. Tübingen: Gunter Narr Verlag.

- Taylor, Christopher. 2003. Multimodal transcription in the analysis, translation and subtitling of Italian films. *The Translator, Special Issue*, 9(2):191-208.
- Taylor, Christopher and Anthony Baldry. 2004. Multimodal concordancing and subtitles with MCA, 2004. In *Corpora and discourse*, ed. Alan Partington, John Morley, and Louann Haarman, 57-70. Bern: Peter Lang.
- Thomas, Jenny. 1995. *Meaning in interaction: An introduction to pragmatics*. London: Longman.
- Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam / Philadelphia: John Benjamins.
- Travis, Catherine. 2006. The natural semantic metalanguage approach to discourse markers. In *Approaches to discourse particles*, ed. Kirsten Fischer, 219-241. Amsterdam: Elsevier.
- Ulrych, Margherita. 1992. *Translating texts: From theory to practice*. Rapallo: Cideb.
- Ulrych, Margherita. 1999a. *Focus on the translator in a multidisciplinary perspective*. Padova: Unipress.
- Ulrych, Margherita, ed. 1999b. *Terminologia della traduzione*. Milano: Hoepli.
- Whitsitt, Sam. 2005. A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10(3):283-305.
- Wichmann, Anne. 2007. Corpora and spoken discourse. In *Corpus linguistics 25 years on*, ed. Roberta Facchinetti, 73-88. Amsterdam / New York: Rodopi.
- Wilson, Deirdre and Dan Sperber. 1993. Linguistic form and relevance. *Lingua* 90:1-25.
- Wynne, Martin. 2004. *Developing linguistic corpora: A guide to good practice*. London: AHDS. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter6.htm>.
- Yang, Li-chiung. 2006. Integrating prosodic and contextual cues in the interpretation of discourse markers. In *Approaches to discourse particles*, ed. Kirsten Fischer, 265-298. Amsterdam: Elsevier.

Additional Online Material

Catwoman movie script. The Daily Script.

<http://www.dailyscript.com/scripts/catwoman.pdf>.

Hoey, Michael. Lexical priming and the properties of text. MonaBaker.com.

<http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesofText.htm>.

Internet Movie Database. <http://www.imdb.com/>.

Linguistic Data Consortium. <http://www.ldc.upenn.edu/About/>.

LDC guide to conventions. Linguistic Data Consortium.

<http://projects.ldc.upenn.edu/SBCSAE/transcription/csaе-conventions.html#ortho>.

Mission: Impossible II movie script. AwesomeFilm.com.

<http://www.awesomofilm.com/script/MI2.html>.

Oxford English Dictionary online. <http://0-dictionary.oed.com.millennium.unicatt.it/>.

PIE: Phrases in English by Fletcher, William. <http://pie.usna.edu>.

Shallow Hal movie script. Drew's Script-O-Rama. http://www.script-o-rama.com/movie_scripts/s/shallow-hal-script-transcript-paltrow.html.

The Devil Wears Prada movie script. The Daily Script.

www.dailyscript.com/scripts/devil_wears_prada.pdf.

AMC Transcribed Movies

- Comar, Jean-Christophe (alias Pitof). 2004. *Catwoman*. Warner Bros.
- Farrelly, Bobby and Farrelly, Peter. 2000. *Me, Myself & Irene*. 20th Century Fox.
- Farrelly, Bobby and Farrelly, Peter. *Shallow Hal*. 20th Century Fox.
- Frankel, David. 2006. *The Devil wears Prada*. 20th Century Fox.
- Roach, Jay. 2000. *Meet the parents*. Universal Studios and DreamWorks.
- Romanek, Mark. 2002. *One hour photo*. Fox Searchlight Pictures.
- Soderbergh, Steven. 2000. *Erin Brockovich*. Universal Pictures and Columbia Pictures.
- Soderbergh, Steven. 2001. *Ocean's eleven*. Warner Bros.
- Van Sant, Gustav. 2000. *Finding Forrester*. Columbia Pictures.
- Wachowsky, Andrew Paul and Wachowsky, Laurence. 2003. *The matrix reloaded*. Warner Bros.
- Woo, J. 2000. *Mission: Impossible II*. Paramount Pictures and United International Pictures.