An ultra-condensed introduction to Discourse Analysis: two-hour tutorial at NICTA (National Information and Communication Technologies Australia)

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso
name.surname@pucv.cl

23 October 2015

Abstract: Linguistics has historically been devoted to the study of language as a system rather than its manifestation in particular texts. In spite of this, Discourse Analysis (DA) has been established for more than 60 years as a place of convergence of many disciplines interested in discourse. This tutorial is aimed at people with little or no knowledge of linguistic terminology and its purpose is to offer a shallow introduction to DA. This tutorial will not specifically address computational procedures for text analysis. On the contrary, the analytical methods explained here are to be conducted manually, but with systematic and almost mechanical methods. This is mainly a practical tutorial and the focus will be on analysing a real text in English. The first part presents the theoretical principles and analytical tools, while the rest is devoted to practical exercises implementing the proposed method. Given the limitations of time, this tutorial only comments upon a short number concepts.

1. Introduction

It is a real pleasure for me to be here today in this not very sunny day in Canberra and I am delighted to have all of you here. As you probably know, I have been entrusted with the impossible task of summarizing 60 years of research in discourse analysis (DA) in a couple of hours. Facing such challenge, what I am going to do is not really to offer a real introduction, because we would need a semester course just to cover the essential readings. Instead, I reckon that the best approach here is to work with a limited set of DA concepts, the basic ones that we would find in any introductory book on the subject.

We will briefly discuss these concepts in the first hour and then do some practice together analysing a short text that we will download from the web, a newspaper article like an opinion piece. We can think of this as a game: I first explain the rules and then we play together. My idea is to project some texts on the whiteboard and then draw on them using these colour markers that have been kindly provided for the occasion. This session is being recorded and I promise (speech act here) that at some point I will produce a slightly improved transcription of what will happen today. As I am not a native English speaker, I would really appreciate it if readers can send me an email to alert about eventual grammar mistakes (or factual inaccuracies, irrespective of the language).

Before we start, I should justify why I pretend to be authorised to talk about this topic. I tend to consider myself a computational linguist, and I came to NICTA for a two-month research stay to collaborate with some of my colleagues here. We are trying to apply machine learning algorithms to corpus-based semantic analysis. My first exposure to DA was some 20 years ago during my undergraduate studies in Argentina, my country of origin. After finishing my degree, I went to Barcelona to pursue a PhD in computational linguistics at Pompeu Fabra University. This is where Teun van Dijk was teaching, and whose work in DA we will discuss in a moment. My intention there was to embark on a project in computational discourse analysis, mainly to implement van Dijk's DA model in a computer program. This was 2003 and I didn't know much about computational linguistics at the time, and less about programming languages. As I started to discover what computational linguistics actually is, I diverted from DA into corpus-based lexical analysis and quantitative corpus

linguistics, another chapter in the discipline. Thus, I ended up doing a PhD thesis on a rather different topic, which is computational terminography. Now I am back doing research in DA at the Catholic University of Valparaíso, in the City of Viña del Mar (Chile). Slowly but steadily, I am in the process of integrating computational linguistic tools into DA.

So, where to start this introductory session on DA? Perhaps the best way is to contextualise it historically. It started as a subfield of linguistics, and most people currently doing DA started as linguists in what was also called Text Linguistics in the sixties and seventies. Thereafter, DA became a space of convergence of many different disciplines, most of them from the realm of social sciences but not exclusively. There has been only marginal interest in DA from the computational linguistics and artificial intelligence communities, in very specific areas such as text generation and text summarisation. I have also not yet seen any real integration between corpus linguistics and DA. They are still kept separated as rather different disciplines, although there have been some attempts to integrate them such as Michael Stubbs' (1996) and Paul Baker's (2006) introductory books. I know that the audience here is mainly composed of machine learning specialists, and I hope that after this tutorial you will see some possibilities of application of machine learning in DA. Again, I have not yet seen any intersection between machine learning and DA. Of course, there have been applications of machine learning algorithms to different subtasks of computational linguistics such as parsing, tagging and disambiguating, but it feels that much more interesting things are about happen in the near future.

After presenting a brief sketch of DA history, I will then walk you through some of the most frequently encountered concepts in DA textbooks, such as the very same concept of text or discourse and the difference between sentence and utterance, cohesion and coherence, macrostructure and superstructure theme and rheme as well as the concept of discourse referents, correference resolution, deixis, discourse markers, modality, intertextuality and textual polyphony. Another reason to choose these topics is because they can be applied to the empirical analysis of a text, which is what we will be doing after the initial discussion of such topics. The final section includes a short list of selected bibliography in English, French and Spanish.

2. Definition and objective of discourse analysis; historical precedents and constitution as a discipline

As said earlier, perhaps the best way to understand DA is to see first what it is not, specifically in contrast with the field of general linguistics. Modern linguistics, defined by the seminal work of Ferdinand de Saussure simply as the science of language, started with a layout of the object of study as language seen as a system (*langue*), i.e., not the particular languages of the world but language as a general abstraction. Opposed to language as *langue*, we can also find language as *parole*, which is the myriad of particular instances –i.e. utterances– of a particular language. It was of course necessary for de Saussure at this early stage of linguistics to try to delimit the object of study with this methodological fiction, as he did with other abstractions. His theory is a system of oppositions, where for instance the synchronic study of language was favoured against the diachronic (historic) point of view, which was the most common at the time.

Linguistics continued its own evolution towards the 20th Century and established its different areas of specialisation such as phonology, morphology, syntax and semantics. However, a series of transformations started to take place mainly outside linguistics, which would put pressure on linguists to address discourse as an object of study. Different authors (De Beaugrande & Dressler, 1981; Lozano et al., 1989; van Dijk, 1978; 1980) suggest that the first precedents of DA are older than linguistics,

mainly in the classical studies of rhetoric and stylistics. The modern history of DA probably began in the early 20th Century with the work of Russian semioticians, anthropologists and literary scholars, among whom were well known figures such as Vladimir Propp, who analysed folk tales, and his colleague Mikhail Bakhtin, who was a pioneer in the study of genre and text typology. Bakhtin is also among the first to propose a notion of a text as a speech act and defined many of the concepts we still use today.

Many contributions to DA came form outside linguistics, in the realm of social sciences such as anthropology, sociology and semiotics particularly in France, with intellectuals such as Claude Lévi-Strauss and Algirdas Julien Greimas. In the UK, however, a great revolution was taking place in the field of philosophy of language, with John Austin and John Searle, who also contributed to establish Pragmatics as a new linguistic subdiscipline. They changed the point of view of language as a system and focussed the attention on its realisation in daily communication. This is when the notions of "speech acts" and "performativity" became ubiquitous in linguistic research: people realised that discourse does things, i.e. that there are things that you can only do with words, such as to apologise, forgive, insult, threaten or declare someone guilty or married. These concepts would be further articulated with the notion of genre and discourse typology (Loureda, 2003). Sociologists also started doing DA on their own when they began to develop the study of social interaction and conversation analysis, called microsociology at the time, with prominent scholars such as Erving Goffman, Dell Hymes or Harold Garfinkel.

There were also a series of transformations taking place inside linguistics. In 1952, the great linguist Zellig Harris used the term "discourse analysis" for the first time in a paper in which linguistic analysis transcended the boundary of the sentence. Other American linguists such as Kenneth Pike and Robert Longacre were also extending the object of study of linguistics in order to better adapt it to anthropological research. In parallel, the French linguist Émile Benveniste started another very influential line of research called the "enunciation theory", and by the early seventies, linguists were already looking at full texts as the object of study.

Around this time is when the Dutch linguist Teun van Dijk started his career. It can be said that he is in the centre of this discipline and almost every text on DA refers to his work. Through the seventies, he and many of his German colleagues were responsible for what is often considered a revolution in linguistics with the development, first, of the field of "Text grammar", later renamed as "Text Linguistics" and finally as "Discourse Analysis". It is undoubtedly with van Dijk that DA finally started as a unified theory, and most of the concepts we will study today are taken from his books.

Of course, this very brief history of the precedents of DA cannot do justice to the discipline, and I have not mentioned the work of interesting people such as János Petöfi, Wolfgang Dressler, Eugenio Coseriu and may others. The number of authors and publications of the field is staggering, and I cannot devote too much time to this introduction. As van Dijk (2011) points out, there are more than 5000 books and many more papers with the words "discourse analysis" in the title or in the abstract.

As for the definition of DA and its object of study, by now it should be clear that, as opposed to general linguistics where the object was language as a system (as a virtuality or abstraction), DA is focussed on real and concrete productions of language. And while earlier linguistic theory had introspection as their only method for data collection, DA instead uses observation and empirical analysis.

Up to now, I have only offered a general picture. Hereafter, I will continue by discussing some specific concepts, with the risk of making this account look rather like a small glossary. But it is the

terminology that we need to understand in order to move forward in our analysis.

3. Basic DA concepts

3.1. Text, or discourse

We need to start with the definition of text. For our purposes, we will not distinguish between the terms "text" and "discourse" and thus we will treat them as synonyms. When we talk about text or discourse as our object of study we refer to oral or written text, although most often we will be working with written material. Oral text is important too, but it is of interest mostly in areas mentioned earlier, such as conversational analysis, the study of speech acts and social interaction in microsociology.

When analysing a text, we will not be interested in its content. We will not participate as readers in the sense that we will not play the game that the text is proposing us. We can say that we are inflicting some kind of violence to the text, or at least a use that was not expected from us. The figure of the author as a person is also not relevant for DA and there is no judgement of the content or the opinion of the author. It is true that we need to have some degree of knowledge of the situational or historical context of a text in order to understand it, but we would not, for instance, try to explain the motivations of the author to say one thing or the other.

In DA we analyse a text as if it was a machine. We just analyse how it is that a text works, thus we are trying to do some sort of reverse engineering on a technology that is unknown to us; not a human technology in the sense that it is not the product of conscious effort like the artefacts of human invention. Here I share the view of Umberto Eco (1984) of the text as a mechanism or natural machine that has its own rules and we are trying to understand how is it that they operate.

Enrique Bernárdez (1982) points out that there are features that distinguish texts of any nature. A text, to be considered as such, needs to have the following characteristics:

- It must be a communicative unit: it has to be recognised as a product of a linguistic system.
- It has a purpose: always, every text has an intention.
- It has an internal structure, which will depend on its genre.
- It is always coherent, or at least we expect it to be.
- It has semantic closure: a text has a beginning and an end. It has to be clear where the boundaries of the text are. In oral speech, boundaries are indicated by a pause or by turn-taking. In the case of written or printed text, the clues are in the layout.
- A text has a context with some participants and it is the product of social activity, which means that we cannot totally dismiss the social context. At some point we will have to be aware of where the text is circulating, especially if it is a reaction or a response to other texts.

3.2. Sentence and utterance

Another distinction we need to learn about is the difference between sentence and utterance. All linguistics, including traditional grammar, sctructuralism and generativism, had the sentence as the limit of their object of study. In DA, in turn, we rather talk about utterances. These are realisations or occurrences of sentences. The distinction is analogue to that of *type* and *token* in general linguistics. The type is the abstract idea of a sign and the token is its instantiation in discourse. The distinction is also analogue to that between meaning and sense. A sentence has meaning, but an utterance has sense.

Sometimes, the same sentence may imply different utterances and vice versa. Consider the following examples:

(1) “I do not know where the Cathedral is”.

(2) “Where is the cathedral?”

(3) “Could you please tell where the cathedral is?”

(4) “I would like to go to the cathedral, but it seems I got lost.”

(5) “Do you want to open the door?”

The meaning of example (1) is the underlying content or proposition that we can easily understand based on our knowledge of English. Other sentences having different literal meaning may however convey the same sense, as shown by examples 2-4. Conversely, example (5), with its unique literal meaning, could have different senses depending on the situation. It could be interpreted as an order, an offer of help or even a request no to open the door (Casado Velarde, 1993).

The meaning of the sentence is encoded in the linguistic system, and if we know the language, we will have no problems to understand it. But to make sense for us, we need to figure out what the purpose of the author is with this utterance in this particular context.

3.3. Cohesion and coherence

The terminology has been changing in the evolving history of DA, but the terms “cohesion” and “coherence” seem to have been established in the field at least since the work of de Beaugrande & Dressler (1981). The difference between cohesion and coherence is analogue to that between syntax and semantics. A cohesive text is one that is grammatically correct and its sentences are well connected. A coherent text, in turn, is one that makes sense. We can imagine a coherent text that is not cohesive, and vice versa.

The notion of coherence is not easy to formalise. The notion of cohesion might be easier because of the existence of grammars. A grammar tells us if something is wrong with a sentence or if there is lack of connection or agreement between sentences. But when is a text coherent? The difficulty is that no text grammar can algorithmically decide if a text is coherent or not. There are, however, some clues. Greimas (1996) proposed the notion of isotopy, defined as a beam of lexical recurrence along the text. In every coherent text we will always find a group of lexical units that are semantically related. For illustration, imagine that we produce some kind of pseudotext, say by retrieving sequences of two or three words that are frequent in the language and we connect them with each other in such a way that the transition from one word to the next would sound normal. The result, of course, will most probably not make any sense. Thus, one way to algorithmically decide between a real text and a pseudotext would be to measure its isotopy, as in every real text there will be a progression of the same topics or referents. This might occur because the same lexical units are repeated in the text or they are different but semantically related, either because they are related in the linguistic system or because this particular text establishes a perhaps new conceptual connection between those lexical units. It should be noticed that the concept of isotopy has been rediscovered a number of times, each time receiving different terms (e.g. “lexical chains” by Morris & Hirst, 1991).

3.4. Macrostructure and superstructure

If we continue with the analogy used earlier when discussing cohesion and coherence as the distinction between syntax and semantics, we can say then that every text is built with a superstructure and a macrostructure, and relate the first to syntax and the latter to semantics. Regarding the macrostructure, it explains the fact that, after reading a text, normal people tend to remember the content but not the exact wording. The macrostructure thus consists of the main propositions that summarise the content of the text. It has been empirically determined by Walter Kintch and Teun van Dijk that given two texts of the same length, it takes longer for people to process the one with a more complex macrostructure (Brown & Yule, 1983).

The term “proposition” is used here in the logical sense, and it can be rendered as the idea that is conveyed in a sentence with independence of how it is phrased. We can use a logical notation to represent a proposition as a predicate-argument structure. Thus, the proposition underling the sentence “Peter gives a book to Mary” could be: buy(Peter, book, Mary). The same proposition could be rephrased as “Mary received a book from Peter” or “This book was given to Mary by Peter” and so on. The macrostructure is thus made of propositions and macropropositions, i.e. with connectives between propositions (we will discuss connectives later in the section devoted to discourse markers).

How our brain builds such macrostructures from a text is far from clear, but van Dijk (1977) has proposed a theory to explain the process. According to him, to build a macrostructure we apply a series of “macrorrules”, consisting of four consecutive steps: deletion, selection, generalization and construction. He explains that when we read a text we delete some parts of it and we select others that are considered relevant; we also generalise skipping details and finally we construct the summary maybe adding elements that were not initially in the text but come from our knowledge of the topic.

The superstructure, which is in turn related to syntax according to the analogy with general linguistic I used earlier, could also be described using another metaphor, as the skeleton of the text (and then the macrostructure would be the “flesh”). The superstructure is also related to the idea of genre or type of text. Texts pertaining to the same genre will have a similar superstructure, while the content (the macrostructure) will differ. Narrative, in general, has the same structure: introduction, complication and end. Consider the classical children story: it has a similar formulae for different stories. In the case of argumentative discourse, where the author is trying to persuade us to believe or to do something, we normally have a different type of superstructure, consisting of premises and conclusions. And again, the same superstructure of an argumentative text may convey different argumentations or macrostructures.

3.5. Theme and rheme

Every text will present the two aspects of theme and rheme, i.e., it will contain something that is already known or that the author assumes that the reader knows, and also something that is new, or that the author presumes the reader ignores. The theme is the selection of the topic, where the author states what is going to be discussed. The rheme in turn is what the author states about such topic.

In a very simple utterance like (6), Randy is the theme, i.e. the given, while the rheme is the information about what he did. In (7), instead, Randy is the rheme, i.e. the new information, while the fact that your car was stolen is already known.

(6) “Randy took your car”.

(7) “It was Randy who took your car”.

Again, we will find this distinction with different terminology in the literature. For instance, van Dijk uses the terms “topic” and “comment”, as used by the American linguist Charles Hockett when analysing sentences. Other terms that have been used roughly for the same concept is the pair “new” and “given”.

3.6. Discourse referents

We know that a text will speak about a limited number of referents, also called “actants”. Referents are the main objects mentioned in the text and can be entities of real or imaginary worlds. Of course, texts can be more or less complex but, on average, we can expect them to select only a limited set from the world of possible referents.

We expect referents to be designated by nouns or proper names, and to be agents, topics, places or some other type of entity. This is another idea that has been inherited from sentence-based linguistics: the notion of predicate-argument structures. As earlier, when discussing the concept of proposition, in the example “Peter buys a book to Mary”, we can distinguish between the actants *Peter*, *book* and *Mary* and the predicate *to buy*, the action taking place.

This does not mean that every noun in the text is going to be a referent. Moreover, and although nouns are the prototypical grammatical category for entities, not all nouns denote entities. Many actions, processes or events can also be denoted by nouns. Verbs can also be the subject of a sentence as gerunds or infinitives, as in “To buy/Buying books compulsively is bad for your economy”. In fact, whole predicate-argument structures can be the subject or object of another such structure, and an action can be denoted with a noun by a process of nominalisation or with a synonym, as in “purchase” or “acquisition” in the case of “to buy”.

Despite these rather complex lexico-syntactic issues, we can relate the predicate-argument distinction and the idea of entity to the much older philosophical notion of substance, which is conditioned by how we humans see the objects in our environment. As far as we know, this table we have now in front of us is more like a process or an event rather than something solid or substantive, as there are lots of things going on at the atomic level. We could even force our language to say this is actually “tabling”, as the action of “being a table”. From that point of view, there is no such thing as substance, or it is an epiphenomenon. But language is conditioned by the way we perceive things (or maybe it is the other way around, there is no way to know for sure). What we do know is that, normally, we perceive a table as an object or entity, and we denote it using a noun.

In short, it can be said that a text selects a number of entities of some sort to be the main referents of the text, and they are normally denoted by nouns or noun phrases.

3.7. Correference resolution

Good authors tend to avoid repetition in their texts, but if they talk about a limited number of referents, then they will have to repeat the name of such referents or at least to mention them by using some other linguistic mechanism like a pronoun. This is what we call correference, or the mechanisms of endophoric relations. I mean by that relations between elements inside the same text, as opposed to

exophoric relations, which are the relations between elements of the text and elements of the outside world (we will review those in the next section).

Endophoric relations in the text allow the author to maintain the referents of discourse without excessive repetition. This can be done using lexical units such as synonyms or hypernyms, or grammatical units such as particles or pronouns. We will classify endophoric relations as anaphoric or cataphoric depending on their position with respect to the referent they are replacing. An anaphoric relation is found when the author refers back to a referent that has been already mentioned in the text, and a cataphoric relation occurs when the author is anticipating the introduction of the referent. Consider, for illustration, the following examples:

(8) “The baby was crying. He was tired and sleepy”.

(9) “Before she was told the news, Mary was having a great day”.

(10) “Peter forgot the papers in the train. Bad news for Mary”.

(11) “The dog has bitten the postman. These animals are unpredictable”.

(12) “Boby doesn't like the postman. Dogs are like that”.

In example (8), we assume that the pronoun “he” refers back to the baby. Example (9), in turn, would be a case of cataphoric relation, because the pronoun “she” is referring to Mary, even when such name has not yet appeared. Example (10) would be a case of ellipsis (also known as zero-anaphora). The element that has been elided is the fact that Peter lost the papers. A non-elided version which still avoids repetition would include for instance a demonstrative pronoun and a verb, as in “This was bad news for Mary”. Finally, example (11) shows a case where correference is maintained by a hypernym, as the correferent “animals” in the second sentence refers to the dog and not to the postman, and the same happens in example (12) with an “instance of” type of relation.

There is a rule of distance here: the longer the distance between correferents in the text, the larger the probability that there will be a repetition or partial repetition. Repetition can also be needed to avoid referential ambiguity, i.e. when it is difficult to know what a particular mechanism is referring to.

3.8. Deixis

We were talking about endophoric relations, i.e., relations between elements inside the text. Now we will discuss exophoric relations, those between elements of the text and elements in the external context. It is quite normal to encounter relations between elements of the text and elements of the context, i.e. the outside world. The most basic example would be something like (13):

(13) “I would like to have that sandwich over there” (while I point with my finger).

This is a typical case of deixis. Deixis is a Greek word meaning “to show” or “to point” something or somewhere. The term is not exclusive of DA; linguists have long been using it. Among those who first noticed this type of mechanism were linguists Otto Jespersen and Roman Jakobson. Deixis became then a central part of the enunciation theory of Émile Benveniste, later continued by Oswald Ducrot and Catherine Kerbrat-Orecchioni. Again, there are many terms to denote these mechanisms, such as *shifters* or *embrayeurs*.

Exophoric relations connect elements in the text with elements in the context of enunciation of such text. These relations can be of a personal, spatial or temporal nature (me, here and now). Every text has these three dimensions or tries to hide them, which is what we typically do when writing a scientific paper, for instance. We try to hide the first person to sound more objective.

Sometimes a text cannot be interpreted without knowledge of the context (the time, place and participants of the interaction). For instance, if I find a note in the floor in the middle of the street, saying “We will meet here tomorrow”, then I would not be able to determine who is going to meet, where nor when. The only way to know it would be to be aware of the circumstances in which the enunciation took place (who wrote such note, when and where).

Deixis is sometimes a source of misunderstanding. For example, now that I am in Australia and when I have to interchange email correspondence with my colleagues in Chile, when I say “today”, for them it is “tomorrow”, because they live like 15 hours in the past. Another source of difficulties is that some grammatical elements may be ambiguous, as they can have both functions, that is, being sometimes a form of deixis and sometimes a form of correference. For instance, demonstrative pronouns such as “this” and “that” can refer to elements pointing somewhere inside or outside the text, as in “this was bad news for Mary” or “this sandwich over here”.

3.9. Discourse markers

Another big topic, very fashionable nowadays, is that of discourse markers (DM). Many scholars are now publishing on the subject, but this did not happen until very recently. DM is also another of those phenomena that received many different terms. They are sometimes called conjunctions (Halliday, 1985); connectives (van Dijk, 1978) or textual markers (see e.g., Fraser, 1999).

DMs are particles that help us find coherence relations in the text. They are considered instructions for the reader on how to connect the different propositions of the text and organise ideas in the author's argumentation. There are many types of discourse markers. Consider, for instance, a very simple example such as (14).

(14) “John is poor but happy”.

Why does (14) work as an utterance? Because normally if I am poor I am then expected to be unhappy, thus the DM “but” here is suppressing such inference. It says something like the following: “what I am going to say next will contradict what you may be expecting from my previous proposition”.

Other markers, in turn, would make a cause-consequence relation explicit, such as in (15):

(15) “The baby was crying. Therefore, the mother picked him up”.

“Therefore” makes explicit that the connection between the two propositions in (15) is of a cause-consequence nature. The same happens in (16). Again, “because” is the DM telling us how to connect the two propositions.

(16) “The mother picked the baby up because he was crying”.

Coherence relations between propositions do not necessary need these connectives, as sometimes they

can be easily inferred by the reader, who will have no problem to see coherence for instance in (17):

(17) “The baby was crying. The mother picked him up.”

Although there is no DM to make the cause-consequence connection between the two propositions explicit, we still can infer that the mother did that because the baby was crying. There are many inferences that we are doing here without being aware. We assume for instance that the mother is the baby's mother, and without any other information, that “him” refers back to the baby. The function of the DMs is then to make it easier for the reader to understand what is being said in the text, but they do not produce coherence relations per se.

Question from the audience (Gabriela): My question is if all languages have this same system of discourse markers. I wonder if there is any language that does not use them.

Reply: I cannot know for sure but I would say this is a universal feature of language.

Comment from another member of the audience (Giulio): If there is the possibility of not using them, then it is conceivable that some language don't.

Reply: I agree in that it is conceivable. In fact, it would be an interesting topic of research to compare the system of DMs of different languages, at least to see if they all use the same types of discourse markers.

DMs also depend heavily on the genre. In formal or academic English (scientific or argumentative discourse) they will be very frequent. But they will not be the same we would find in narrative discourse. The same happens in oral or colloquial English: we will find instances of them but not of the same type.

(18) “Well, you know, when they said they where coming, we were not very happy”.

In (18), fillers such as “Well” and “you know” are considered conversational DMs.

Second language speakers often misuse DMs. We tend to attach to some of them and use them excessively. “Therefore”, “hence” or “thus” are typical examples. You can tell when a paper has been written by a non-native speaker because of the use of DMs.

When linguists began to realise that DMs exist they tried to collect them and organise them in taxonomies. There are different types of DMs according to their function, and they cannot be used freely in every context. The context imposes restrictions on the type of connectives we use (van Dijk, 1978). Consider the following example:

(19) “Mary was playing tennis and Joe was reading a book”.

In that case, the coordinate conjunction “and” would be functioning as DM and is used to add new information. But a condition of felicity here is that the two propositions are semantically compatible. In turn, we cannot say something like (20):

(20) ? “Mary was playing tennis and Joe was an engineer”.

DMs can pertain to different syntactic categories, such as conjunctions, adverbs, prepositional phrases, idioms, and so on. They seem to be outside of the syntactic structure of the sentence and they do not participate directly in the sentence's propositional content. If we project the syntactic tree of the sentence, we would never find them playing a role, they are always outside. They instead affect the whole sentence or the relation between the sentence and other chunks of text. If we see that they are indeed playing a syntactic role, then they may just not be DMs.

The literature says they appear normally at the beginning of the sentence, but if we inspect how they are really used in a large corpus of Spanish, we will find that this is not always the case. They appear sometimes in the middle, frequently between comas.

There is no single taxonomy of discourse markers yet. There are different taxonomies, and it is to be expected that they will converge into a single standard, but we will still have to wait for that. None knows yet how many there are. Some typologies have been attempted based on the difference between paratactic (coordinate conjunctions) and hypotactic (subordinate conjunctions) classes or depending on the type of relation they express between propositions. We can also try to classify them by their function. Here I just mention some of the most frequent categories and some examples based on the taxonomy offered by Martín Zorraquino & Portolés (1999).

A first large class is the structuring type, used to organise the ideas in the text. They are used to comment, to enumerate or present ideas in sequence or to digress or present an idea that does not strictly follows from the line of thought presented in the text.

-Structuring:
- -Comment: *Well, Indeed*
- -Organize: *first this, then that, on the one hand, on the other, finally*
- -Digress: *by the way*

Another class consists of the connectives. Their prototypical function is to connect ideas in the text, to add something new, to make a cause-consequence relation explicit or to present a counter argument, i.e. to suppress some inference that could follow from a previous proposition.

-Connection:
- -Addition: *also, moreover, furthermore, in the same way*
- -Consecutive: *therefore, consequently*
- -Counter argumentation: *but, however, on the contrary, nonetheless*

Another class consists of those markers used to reformulate, maybe to rephrase something we said in order to explain it better, or to correct or rectify it. We also dismiss or distance ourselves from some point of view we presented in the text or summarise what we said, typically at the end of the text.

-Reformulation:
- -Explanation: *that is, in fact, in other words*
- -Rectification: *better said*
- -Dismissal: *anyway, in any case*
- -Summarisation: *in the end, in conclusion*

Another class is composed of those DMs used for reinforcement. In argumentative discourse it is very frequent to use expressions to intensify or reinforce what we are saying.

-Reinforcement: *actually, clearly, above all*

Finally, another class consists of those we use to exemplify when we present abstract or complex ideas.

-Exemplification: *in particular, for example, for instance*

I insist that there are many more DM typologies, but these types are probably the most common.

3.10. Intertextuality and textual polyphony

Every text is always related to other texts. A text is just a node in a very wide network of previous and future texts. Texts are always referring to other texts, either explicitly or implicitly. We owe this vision to M. Bakhtin, who first decomposed the idea of author as a unique individual into a myriad of different "enunciators". This is the meaning of the term "polyphony": the idea that there are many voices in a text, and not just the one of the author.

An author will use the words of other authors to dispute them or to use them as a premise to support his or her own ideas. The source of these ideas can be clearly asserted, as it is usually done in a scientific text, or not, but still one can tell that they refer to other texts. In some cases the source of a point of view is not explicitly attributed to a specific text, as in "Many people believe that this is time for action".

3.11. Modality

The last topic in our discussion is Modality. This is a very old subject which has been studied for a long time in medieval logics. With the terms "dictum" and "modus", philosophers separated the content of a text from the point of view or attitude from which it is narrated, i.e., the author's position concerning what he or she is communicating. In classical grammar, modality instead means for a sentence to be expressed in declarative, interrogative, imperative or exclamation mode. In modern DA, however, we analyse expressions of a text that show the attitudes of the writer, which are sometimes unconsciously revealed. These clues are found in a metacommunicative level and indicate the certainty, evaluation, beliefs and inclinations of the author with regard to the main message or propositional content. We can say it is a proposition (modus) about another proposition (dictum).

As I did with DMs, here again I will consider only some of the most frequent types. Take, for instance, a proposition such as (21):

(21) "It is going to rain this afternoon".

There is no such thing as a text with zero-modality, in the sense that there is always some type of modality in every text. That said, we may also say that (21) conveys the most basic meaning or the essential information. We may, however, rephrase the same content this time expressing it with a layer of epistemic modality, as in (22):

(22) "I think it will rain this afternoon".

An epistemic modality may express certainty or lack of. Again, I am saying something about what I am saying, and that is why we say it is a proposition on top of the other.

In the same vein, one may say something and at the same time express a desire (or lack of). This would be the case of the volitional (or volitive) modality, as in (23) or (24):

(23) “How I wish it would rain today”.

(24) “I hope it doesn't rain today”.

Another type is called the deontic modality, and it occurs when an author expresses that something has to be done or it has to be believed, as in (25) or (26):

(25) “It is crucial that our political leaders take action now”.

(26) “You should not smoke in front of your children”.

We will consider also the evaluative or axiological modality, as it is used to add positive or negative values to what we say, exemplified in (27):

(27) “Unfortunately it is going to rain today”.

Another frequent type is the veridictory modality. This one is often presented when the author claims that something is true or false, or it is a lie or a secret. It is usually found in negations, as in (28).

(28) “It is not true that rain will fall today”.

Finally, another form of modality is called alethic, when one asserts the possibility or impossibility of something, as in (29).

(29) “It will possibly rain today”.

It is sometimes very difficult to distinguish one type of modality from the other, especially in the case of the epistemic and the alethic modalities, because something is possible or impossible always according to someone knowledge of some facts.

When students learn to write academic papers, it is important that they know when they are using modality. Ideally, a scientific text should not use any kind of modality, because a scientist should not say what he or she likes or dislikes, or what should or should not be done or believed. Likewise, scientists are also not expected to express a variable degree of certainty on what they are saying, and so on. Again, as I said before, there is no text with zero-modality. The very fact that an author decides to talk about something is already some expression of modality because it is a way to state preference for some topic over another.

4. Time to get our hands dirty

We are now going to select a text from today's news websites. The type of text I think is most appropriate for this exercise is the opinion piece, because there we have an author trying to convince us of something we have to believe or do. Scientific abstracts are also a good type of material to analyse, but they might be more complicated, so the opinion piece type will be our best option now.

We could select any text from the BBC or the New York Times, but I was thinking about the Deutsche Welle, because they write in English in a simple style, and it is also interesting because of their rhetoric. In a way, they represent the Public Relations office of the German Government, so they try to improve the image of their Government and thus it is easy to find typical argumentative structures. If we visit today's edition of the Deutsche Welle's website (<http://www.dw.de/>) we might find something useful such as this (<http://www.dw.com/en/opinion-steinmeiers-mid-east-trip-a-tough-sell/a-18794161>):

Steinmeier's Mid-East trip a tough sell
by Engel Dagmar

The deepest place on the earth's surface is the Dead Sea. A comparison to the political situation in the Middle East begs to be made. German Foreign Minister Frank-Walter Steinmeier used the analogy during the last leg of his diplomatic trip to the region: The situation is sinister, he said, and it is especially grim in Syria.

Two regional powers could do much to see that the situation changes: Iran and Saudi Arabia - neither a shining example for democracy, nor the defense of human rights. These nations despise each other so much, direct commercial flights between them do not even exist. Considering Steinmeier visited them one after the other, one can only assume that he is endeavoring to mediate; that he is trying hard to get Tehran and Riyadh to come to the negotiating table and work out a solution for the Syrian crisis.

But one of the most repeated lines that Steinmeier used in his press statements was: "We are not here to mediate." When asked, why are you here then? He says: "To figure out what might be possible." To determine where bridges might be built and pass along that knowledge to those that are responsible for mediation. But who would that be? "Staffan de Mistura, the UN Special Envoy for Syria." But how much backing does he have?

It is certainly true that Germany is not a major player in the region. And that the German army is not the tool with which to leverage negotiations. Therefore, Germany is well advised to operate in concert with its partners in the European Union on foreign policy matters. And indeed, it has enough to do in terms of its mediations in Ukraine. Ultimately, the USA and Russia are pursuing their own interests in the Middle East - and with a lot more force.

Nevertheless: When the German foreign minister says at the end of his journey that a deciding factor in solving the crisis will be that states and personalities will have to step forward to accept risks and responsibilities, it applies to him as well.

Mediating takes time. A solution for the Syrian crisis will take time, and people will continue to flee by the thousands. A million will come to Germany this year alone. In such situations, the mediator can quickly be declared a guilty party. A mediator can fail.

So let's call it something else, let's call it brokering, instead of mediating; after all, Germany enjoys the reputation of being an honest broker in the region. We can go along with the foreign minister and call his activities soundings, bridge building, or holding talks.

I would like to call it action.

To analyse this text, we will proceed following the steps from the point of the selection of the referents. Let's imagine that we are a computer and we will try to think and act as one, very systematically, mechanically, if you want. Our first attempt will be to detect the main referents of the text. This text selects a number of topics, or actants. In this case it is easy to identify them: Frank-Walter Steinmeier, Germany, the Middle East, Syria, Iran and Saudi Arabia. And I think the last two can be considered a

single actant, albeit they are presented as opposite players in the text. Other referents of secondary importance, considering how often they appear in the text, could be the UN Special Envoy, the European Union, USA, Russia and Ukraine.

Computationally, it would be possible to detect the main referents of the text by calculating the frequency of occurrence of nouns and noun phrases. But to do that we need to proceed with anaphora resolution, because that helps to increase the frequency counts of the terms.

Let us now analyse how each of these referents are maintained through this text. We will try to find the correference relations in the text. Let us begin by Steinmeier. Pay attention at how he is presented, at line 5, as “German Foreign Minister”. This is a typical construction. Umberto Eco, who is also a very important figure in discourse analysis, stated that an author will always have a model of the reader. The author does not know for sure if we, as readers, know who Steinmeier is. She knows that some of us know Mr. Steinmeier while some others may not. So she needs to inform those who don't know who is the Minister but also she does not want to look like she is assuming that we don't know. She cannot state that explicitly by saying something like “Steinmeier is the German Foreign Minister”. So this is a typical way to avoid loosing those readers who don't remember who Steinmeier is and at the same time not assume too much about the reader's knowledge of German politics. So this would be the first mention of the referent “Steinmeier” in the text. Now let us see what correference mechanisms we can find in this text regarding this particular referent.

We find a first instance of an anaphoric relation. The author does not want to repeat the name of this Minister, so she uses a pronoun, at line 6: “he said”. We call it an anaphoric relation because it goes back to the referent.

There is another instance of anaphoric relation but it is a case of partial repetition at line 12 (“Steinmeier”). Then, again a pronoun at the same line and at the next one (“he”); partial repetition again at line 16; the pronoun again at line 17. Again “German foreign minister” at line 28 as well as the possessive pronoun “his” and another pronoun “to him” at line 30. “The mediator”, at line 33, is also referring to Steinmeier and then we find “the foreign minister” at line 37 and “his” at line 38.

We will now leave Steinmeier there and continue with the rest of the referents. Iran and Saudi Arabia. Let's see if we can keep them as a single referent.

Gabriela: “Two regional powers”, at the beginning of line 9, would be a cataphoric relation of Iran and Saudi Arabia.

Rogelio: Exactly. And then we have “neither” at line 10, as a pronoun, establishing an anaphoric relation, same as “These nations”, in an instance-of type of relation, also conducting a correference relation. The same occurs with “each other” (line 11).

Giulio: There is an instance of “one after the other”, at the beginning of line 12. These are also Iran and Saudi Arabia. And in “direct flights between them” at line 11 towards the middle, “them” refers also to Iran and Saudi Arabia.

Linda: And then we have the capital cities, Teheran and Riad, at line 13, also in a correference relation.

Gabriela: What kind of correference would that be?

Rogelio: There are different types of classifications. One classification is based on the position of the mechanism with respect to the referent and it determines if it is anaphoric or cataphoric. But then we have the type of mechanism that is being used. Typically we will see pronouns, synonyms and hypernyms used as correference mechanisms. In this case, the capital city would be a part-whole type of relation. This is a figure of speech called synecdoche, typically seen in poetry but also in daily speech. It is like when you say “head of cattle”. You usually do not refer just to the head of the cow, but to the whole animal. A similar case would be the metonymy, when you say for instance “the chair of the institute”. You usually do not refer to the chair but the person sitting on that chair, i.e. the one with the highest authority in the institution. And yes, these are mechanisms that can be used to establish a correference relation.

Linda: What about “The region” at line 22?

Rogelio: I think it would refer to the Middle East, another of our referents which is different from Iran and Saudi Arabia, because it would be the zone of conflict. It appeared in the title and at line 5, and it is repeated at line 26.

Giulio: And there other instances of “the region”, at lines 6 and 37.

Rogelio: I think in this text Syria and the Middle East are presented as the same referent. Syria as such appears at lines 7 and 20. It appears as “the Syrian crisis”, at lines 14 and 32, and probably “the crisis”, at line 29, is also establishing a correference relation as a paraphrase. Then we have “Germany”, another of our referents, repeated at lines 22, 23, 33 and 36. Also we have the “German army” at line 22 as a part-whole relation and the pronoun “it” referring back to Germany at line 24. Maybe, we could have considered Germany and Steinmeier as a single actant in this text.

Neil: I wanted to say that I have just tried this same text using the Stanford Core NLP parser (<http://stanfordnlp.github.io/CoreNLP/>). The result of the anaphora resolution is not very good, though.

Gabriela: It would be very difficult for a computer to detect those types of correference relations.

Giulio: If it is difficult to do it by hand, it will take a long time before a computer program can do it.

Rogelio: Indeed, it is not an easy task. But I think it must be possible to do it computationally. The problem with anaphora resolution programs in computational linguistics is that they usually do not identify the referents of the text first. I think here lies the key to solving the problem, because there will always be a limited number of referents in a text. If you have a list of the referents of the text, finding correference relations becomes a problem of classification. If you identify something in the text that must be a correference mechanism like, say, a pronoun, then you just have to check in the list of referents that you have previously identified and then you calculate the most probable one based on different measures such as the distance in words.

So let's get going. We identified the main referents and we identified the correference relations. Our next step would be to identify cases of deixis. I do not think there is much deixis in this text.

Giulio: How do we interpret the “we” towards the end of line 16, in “we are not here to mediate”. Is referring probably to himself as a representative of the Government.

Linda: “We” could be “Germany”.

Rogelio: It could be an inclusive “we” as in “we, all together”, or an exclusive “we”, as in “we, the Government”. It is somewhat ambiguous.

Neil: It also could be “me and my assistants”. He is referring to his role in the conflict.

Giulio: And then, the next line, line 12, says, “Why are you here then”. “You” would refer to Steinmeier.

Rogelio: The problem here is that we are dealing with three different texts. The text of the journalist, the quotation from Steinmeier and then the quotations from the journalists in the press conference. So it gets a little bit complicated. We will ignore then the quotations because we will consider them different texts, texts inside the text.

Linda: What about the instance of “here”, at same line: “Why are you here then”. Where is here?

Rogelio: “Here” is the Middle East. One interpretation is that the journalist is with Steinmeier in the Middle East, specifically in Saudi Arabia, if we interpret that part as the voice of the journalist. In that case we would consider it an indication of spatial deixis. It is though not clear if it is her voice or the voice of other journalists present in the press conference room. Remember, deixis is everything that relates with the context of enunciation of this text. It can be personal pronouns but also time expressions as well as clues about the location.

Giulio: And what about “those”, at lines 18-19. “Those that are responsible for the mediation”.

Rogelio: In that particular case that would not be a deictic but a correference mechanism. I suspect it is referring to Staffan de Mistura, another of the actants. I think in this text there is not much reference to the time, space and person of the author except in the end, in the last two paragraphs with the first person appears. Especially at line 40, when she says “I would like to call it action”. This is a frequent structure. Many argumentative texts begin without presenting the voice of the author. That is, the voice of the author is hidden to present a more objective text and then it appears suddenly towards the end of the text. This produces a rhetorical effect of approach from the author to the reader and so it increases the intensity of the text, not consciously perceived by the reader. It was an objective text and then it becomes personal at the end.

But let's get going. The next level we would have to analyse consists of the discourse markers. So, what do we have here as discourse markers? It is very tricky to find instances of discourse markers except when they are presented in their typical forms. You can compile lists of the most frequent markers, but you will always find new expressions that are used with the same function. And we will also find the same markers used with different functions. We can consider the “and” at line 6, in “and it is especially grim in Syria”, and additive marker. And the “but”, at line 16, a counter-argument marker. But the two “buts” at lines 19 and 20 are instead conversational markers: they are not meant to present a counter-argument.

Linda: There is a “therefore” at line 23.

Rogelio: That would be a cause-consequence type of discourse marker. Also the “and” at line 22 as well as the next one at line 24 can be considered additive markers.

Giulio: What about the “and” at line 13 where it says “to come to the negotiating table and work out a solution for the Syrian crisis”? Isn't that a discourse marker?

Rogelio: Not in that particular case, I would say.

Linda: So not every occurrence of the conjunction “and” is a discourse marker. Then what are the rules to tell, especially if we are supposed to do this computationally?

Rogelio: It is a matter of scope. If it connects two different propositions, then you can consider it a discourse marker. But consider, for instance, the following example: “Mary loves cats and dogs and Peter loves to drink beer”. The first “and” would not be a discourse marker, because it is part of a single proposition. The scope of that “and” is not affecting two different propositions. Thus the scope of this “and” is not connecting different parts of the text. But it is different with the second, which is indeed connecting two different propositions, and therefore would be a case of an additive marker. Now, regarding the rules on how to distinguish them computationally, they will depend on the syntactic parse tree of the sentence. There are typical behaviours of discourse markers, as I said before. They can appear at the beginning of a sentence or in the middle, most frequently separated by commas. But this is not always the case, and this is why it is so difficult to determine when a conjunction is working as a discourse marker. If it is difficult to do it manually, more difficult will be to do it computationally. There are some cases that we can recognise immediately, and there are others that produce much more confusion. One method to do this is to replace it with another marker that would have the same function. If we can do that without altering the sense of the text, then we have spotted the right marker. For instance, I could replace the instance of “And indeed” at line 24 with “moreover” or the “nevertheless” at line 28 with “however” and the general sense is not altered, thus they have the same function and therefore they can be considered as equivalent markers. This would be the commutation or permutation method. Other markers we can find in this text are “Ultimately”, at line 25, for a summarisation or recapitulative marker, as well as the “and” at lines 26 and 32, as additive markers. The “So” at line 36 would be a conversational marker, more common in oral speech than it is in written text, and finally the “after all” at line 36, again with a recapitulative function, or most probably used as a justification.

Let's move on now and try to find instances of modality, and I am afraid that is all we will have time to do today. When we find expressions such as “The situation is sinister” (line 6) and “especially grim” (line 7), we have evaluative statements and therefore we see axiological modality.

Giulio: What about “could do much”, at line 9?

Rogelio: That would be an instance of the alethic modality, because the author is stating that something is possible or not possible.

Neil: I think it's not alethic, it's a statement of belief.

Rogelio: Again, it is very difficult at some points to distinguish between the epistemic and the alethic modality.

Giulio: And what about “neither a shining example”, at line 10. Doesn't that convey a judgement?

Rogelio: Yes, that is a judgement indeed. And it is also irony and a figure of speech. But in our small

taxonomy it would be another case of axiological modality. Then we find “One can only assume that”, at line 12. That would be a case of epistemic modality. The author is not giving us this information straight. She says “I believe in this because such and such”.

Giulio: And what about “These nations despise each other”, towards the end of line 10.

Neil: It is not a statement of her own belief, it is a kind of quotation.

Rogelio: Yes, but I guess you could express the same information without using this evaluative vocabulary. There certainly is an evaluative modality here because you could say the same content in a more neutral way. But we don't have to agree! Unfortunately this is not mathematics.

Giulio: “Is trying hard”, at line 13.

Rogelio: Yes, it is possible to relate it to the fact that the journalist thinks that Steinmeir is trying to achieve something that is not possible. It gives the feeling that she is doubting of his possibilities of success.

Giulio: Yes, but how could you say that in a neutral way? How can you make that non-judgmental?

Rogelio: As I said earlier, there is no text with zero-modality. But it is a question of degree. You do not want your students to write papers with much modality on them. You want your students (or your journalists, for that matter) to make their text look neutral and objective. Then we have “It is certainly true”, at line 22. She is very sure about something, that would be a case of epistemic modality. Then “German army is not the tool...”, at the same line. This one would be a case of veridictory modality, the modality related to truth, falsity, lie and secret. This modality is often associated with negations, when the author says that something that is being held as true is actually not true. Finally, and in coincidence in this case with the emergence of the first person in the last two paragraphs (“So let's call it something else, let's call it brokering ”, “We can go along ”, “I would like to”) we have a case of deontic modality, where the author is now trying to persuade us.

Well, we are running out of time and I think we should stop this here and try to wrap up.

5. Final discussion

For the final discussion, I would like to start with the most difficult questions: What is DA good for? What is its purpose? What can we do with it? To be perfectly honest, I am not entirely sure, but I have my opinion. A legitimate purpose could be pedagogical: if you teach your students some DA tools they may start writing better, because they will become aware of the text's mechanisms and learn how to control them. They will for instance be thinking about the correference relations and therefore try to be more clear, to avoid repetition but also to avoid referential ambiguity and produce more coherent texts.

But there are also scientific and/or practical purposes. It is now too early to say that we will be doing computational DA because there are many problems lying ahead. And yet there are many things that we can already do, but that, however, will have to be the topic of another talk: computational analysis of discourse. Such endeavour would be deeply related to information extraction and document categorisation, among other possibilities we have today. In fact, I have already conducted text categorisation using DA tools. My students and I have been successful in algorithmically categorising

documents by genre using the features we have been discussing today.

Another purpose of DA, and a completely different line of research also very popular nowadays, is steered by Prof. Teun van Dijk, who tries to advocate for political activism based on Critical Discourse Analysis (CDA). He and others working in that particular field are trying to fight against injustice, racism and gender inequality by showing how the practices of the evil powers are presented in discourse and the media, and then exposing and disarming the discourse strategies of power, at least to denounce such practices.

Gabriela: And also the analysis of ideology and how it is propagated through the media.

Rogelio: Yes, showing the strategies and the fact that texts are never candid. There is always an agenda, an intention, concealed or not. As Bakhtin said, language is the arena of continuous class struggle.

Jaume: Is it now a standard practice to use computers for some of the analyses in DA?

Rogelio: Yes, for some of the analyses. I can think of the study of discourse markers, for instance. But in that case the techniques are closer to standard practices in corpus linguistics, where normally you do not analyse a single text but thousands of them. Imagine, for instance, that you want to study the syntactic behaviour of a marker such as “however”, or to see if it is always used for the same purpose or function. What you would do then is to look up such unit in a large corpus using a concordancer, for example. That will give you thousands of lines with contexts of occurrence of such unit in many documents. This is something we did not had time to cover in this tutorial and should be the topic of a different talk. There are many tools for corpus linguistics, and many concordancers that you can download for free or can even be used online, such as Webcorp (<http://www.webcorp.org.uk/>) or Mark Davies' software (<http://corpus.byu.edu/>). Concordance extraction is the most basic operation in corpus analysis. There are many other more complex techniques.

Chris: I can imagine how these tools could be used for instance to compare the use of discourse markers in different languages.

Rogelio: Exactly. In that case you would use a different technique also from corpus linguistics which is parallel corpus processing. Another popular example on the web for that kind of approach is Linguee (<http://www.linguee.com/>). You have there a corpus of texts alongside their translation to another language. The alignment is automatic and sometimes there are errors, but most of the times it is correct. There we can see for instance which is the most frequent translation of a discourse marker or other kinds of units. Of course this is a tool intended to be used manually, but if we process a parallel corpus computationally, we can compute some co-occurrence statistics and come up with the translation of discourse markers automatically. The same method can be used to find synonyms in a language, as they are presented as different translations of the same unit in the other language. If we look, for instance, the Spanish expression “Sin embargo”, we will see that equivalent expressions that appear in the translations are “however”, “nevertheless”, “but”, and so on. This way we could organise discourse markers and also match them between two languages. But I would say this is rather classical corpus linguistics and not discourse analysis. As I said earlier, the real application of computers to discourse analysis has not yet began.

Gabriela: In the last CONLL NLP conference the topic was the automatic detection of theme and rheme with machine learning (<http://www.cs.brandeis.edu/~clp/conll15st/>), and there you can see

that they are doing some progress on the field. The results are not very good yet, but they are moving towards that direction. It is the first competition, and I guess they will get better the next time. I imagine that just the annotation by humans must be very costly. It is a small test, but they now have this annotated corpus as a gold-standard that can be used to test computational methods.

Neil: It must be a very difficult task because language has so much ambiguity and in order to solve it you need knowledge outside the text, that is, of the world, and that is very difficult to model.

Rogelio: One answer on how to model the knowledge from the world comes from the field of distributional semantics. It does not provide an explicit model of the knowledge of the world, but an implicit one that is still very useful. It provides semantic relations between words because of the computation of how they co-occur in large corpora. You may not be able to determine the precise nature of that relation, but still you know they are somehow semantically related. For instance, using that technique, you do not need a large database to know that Teheran is the capital of Iran. You can know that implicitly by inspecting a corpus because both words appear very frequently together. That is one way to solve for instance word sense disambiguation.

Giulio: But sometimes ambiguity is not possible to be solved. I remember some example I read once: “I see that petrol can explode”. There are two different interpretations.

Rogelio: Yes, but there is one interpretation that is more plausible depending on the context of occurrence. It is like Chomsky's classical example: “I saw the man with the telescope”. It is a question of probability. You will choose the most probable.

Chris: Is there a standard graph or a way to represent the content of a text?

Rogelio: Yes. There have been different proposals over the years. There is a famous one by Robert de Beaugrande and Wolfgang Dressler from 1981. They proposed a graph representation of the macrostructure of the text, inspired on graphs for the representation of syntactic dependency analysis, which in turn goes back to the work of Lucien Tèsniere, a French structuralist linguist.

Gabriela: Also, Mann & Thomson's Rhetorical Structure theory (RST) can be considered another proposal in that sense.

Rogelio: Yes, unfortunately we did not had time to talk about it. I thought it would be better to have first some basic notions before reading about RST.

Chris: This type of study reminds me of music theory analysis, where you take some piece of music and reduce it to a simple form, and then they can relate different parts of the piece and say, for instance, this part is an elaboration of this other part.

Rogelio: And that reminds me of Zellig Harris notion of discourse analysis, from 1952, of identifying units in the text and then the relation between units as what he called “transformations” of the same idea. Chomsky later used the same term with a completely different meaning. The original idea was that you can transform a sentence into a different surface realisation but with roughly the same sense. The typical example would be the difference between the active and the passive voice. We can think of other examples of transformations for instance in the case of noun phrases. We can express the same term in English as a sequence of noun-noun or as noun-preposition-noun as you respect the hierarchical structure of the original, i.e., the heads continues to be the same. As in “lung cancer” and “cancer of the

lungs”. But this might be a little bit off-topic.

Chris: Is there a mechanical way to compute these transformations?

Rogelio: That was Harris' aim, because he was one of the earliest computational linguists. But in his time he obviously didn't have the computational power we have today. Now this is part of the field of information extraction, which, as you probably know, is a very active field in the present. Essentially, the idea is to find some type of abstraction from the always varying surface of the text.

Now I am afraid our two hours have gone and, as you are busy people, I will let you go back to your work, not before stating again that I am immensely grateful for having the opportunity to discuss these topics with you. Some of you participated with comments and others were very quiet, but I hope you all enjoyed it. Rest assure that if you ever want me to do something like this again, I will not hesitate to jump into a plane and come back to Australia in less than 14 hours.

6. A very short list of bibliographic titles in English, Spanish and French

6.1. Selected bibliography in English

Baker, P. (2006). Using corpora in discourse analysis. London: Continuum.

Brown, G.; Yule, G. (1983). Discourse Analysis. Cambridge: Cambridge University Press.

De Beaugrande, R.A.; Dressler, W.U. (1981).Introduction to Text Linguistics. London: Longman.

Eco, U. (1984). The Role of the Reader: Explorations in the Semiotics of Texts. Indiana University Press.

Fraser, B. (1999). What are discourse markers? Journal of Pragmatics, 31: 931-952.

Halliday, M.A.K. (1985). An introduction to functional grammar. London: Arnold.

Harris, Zellig S. (1952). Discourse Analysis. Language 28(1):1-30.

Morris, J.; Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17(1), 21–48.

Paltridge, B. (2006). Discourse analysis. London: Continuum.

Pons Bordería, S. (2001). Connectives/Discourse markers. An Overview. Quaderns de Filologia. Estudir Literaris. 6:219-243.

Renkema, J. (1993). Discourse studies: and introductory textbook. Amsterdam: John Benjamins.

Sinclair, J. (2004). Trust the text: language, corpus and discourse. London: Routledge.

Stubbs, M. (1996). Text and Corpus Analysis. Oxford: Blackwell.

Van Dijk, T. (1977). Macrostructures. An interdisciplinary study of global structures in discourse, interaction, and cognition. Hillsdale, NJ: Erlbaum.

Van Dijk, T. (2011). Discourse Studies. London: Sage.

6.2. Selected bibliography in French:

Anscombre, J.C.; Ducrot O. (1983). L'argumentation dans la langue. Brussels: Mardaga.

Bajtin, M. (1984). Esthétique de la création verbale. Paris: Gallimard

Ducrot, O. (1980). Le Dire et le Dit. Paris: Minuit.

Greimas, A. (1966). Sémantique structurale : recherche de méthode. Paris: Larousse.

Kerbrat-Orecchioni, Catherine. (1980). L'énonciation de la subjectivité dans le langage. Paris: Armand Colin.

Vignaux, G. (1976). L'Argumentation. Essai d'une logique discursive. Genève: Droz.

6.3. Selected bibliography in Spanish:

Anscombre, J.C.; Ducrot O. (1992). La argumentación en la lengua. Madrid: Gredos.
Bajtin, M. (2005). Estética de la creación verbal. Buenos Aires: Siglo XXI.
Bernárdez, E. (1982). Introducción a la Lingüística del Texto. Madrid: Espasa-Calpe.
Bernárdez E. (1995). Teoría y epistemología del texto. Madrid: Cátadra.
Clasamiglia, H.; Tusón, A. (1999). Las cosas del decir. Barcelona: Ariel.
Casado Velarde, M. (1993). Introducción a la gramática del texto en español. Madrid: Arco Libros.
Ducrot, O. (1986). El decir y lo dicho: polifonía de la enunciación. Barcelona: Paidós.
Eco, U. (2000). Lector in fabula: la cooperación interpretativa en el texto narrativo. Barcelona: Lumen.
Kerbrat-Orecchioni, C. (1997). La enunciación: de la subjetividad en el lenguaje. Buenos Aires: Edicial.
Loureda, O. (2003). Introducción a la tipología textual, Madrid: Arco Libros.
Loureda, O. y Acín, E. (2010). Los estudios sobre marcadores del discurso en español hoy. Madrid: Arco Libros
Lozano, J.; Peña-Marín, C.; Abril, G. (1989). Análisis del discurso: hacia una semiótica de la interacción textual. Madrid: Cátedra.
Martín Zorraquino, M. A. y Portolés, J. (1999). Los marcadores del discurso. En I. Bosque y V. Demonte (eds.) Gramática descriptiva de la lengua española, vol. 2. Madrid: Espasa, pp. 4051-4213.
Portolés, J. (1998). Marcadores del discurso, Barcelona: Ariel.
Van Dijk, T. (1977). Texto y contexto. Madrid: Cátedra.
Van Dijk, T. (1978). La ciencia del texto. Barcelona: Paidós.
Van Dijk, T. (1980). Las estructuras y funciones del discurso. Mexico: Siglo XXI.
Vignaux, G. (1976). La argumentación: ensayo de lógica discursiva. Buenos Aires: Hachette.