ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO di LAUREA IN
MEDIAZIONE LINGUISTICA INTERCULTURALE

ELABORATO FINALE

NATURAL LANGUAGE PROCESSING IN SPEECH THERAPY: AN ITALIAN CASE STUDY

CANDIDATO:
Andrea Grillandi

RELATORE:
Mazzoleni Marco

ANNO ACCADEMICO: 2018-2019

Primo appello

Acknowledgements

I would like to thank my parents and my sister for the constant support they give me.

A warm thank to my thesis supervisor who helped me through this work of mine. I want to thank her for the precious knowledge she shared.

A heartfelt thank to Beatrice, she always stands by my side and helps me.

I would also like to thank my friends, without them life would be extremely boring.

Introduction

Computational linguistics studies language(s) in order to create models which categorize and aim at representing the linguistic phenomena of a given language. Automatic machine translation, speech recognition software, search engines able to analyze the vastest of the corpora we know: the World Wide Web. Natural language processing technologies are everywhere around us and we are getting more and more used to be side by side with them in our day-to-day life.

However, natural language processing does not only come in handy to most of us everyday, it may also serve more specific scopes. It is the case of speech therapy. In my thesis I will show how some simple computational linguistics applications could be very useful in the medical field, in particular in speech therapy.

Not only will I introduce what computational linguistics is and what are some of the challenges it is still facing today, but I will also describe a more specific application of natural language processing: dependency parsing.

The thesis also includes a chapter on Developmental Language Disorder, in which I will describe in detail its main characteristics and the differences in its characterization among different languages.

As a conclusion, I will write about the thesis' Case Study, namely a pilot conducted in Tuscany about children suffering from DLD. In the last chapter I will clarify the link between computational linguistics and developmental language disorder, showing how dependency parsing can be used for the statistical elaboration of clinical studies results. The last chapter will also overview all the software I used to conduct my research and to semi-automatically process linguistic data.

Contents

List of Figures

Chapter 1

Computational linguistics

Linguistics as a science is interested in linguistic phenomena, namely what people say, write, or express through any other media, in various contexts. Computational linguists aims at creating computational models which categorize and represent the linguistic phenomena in which linguists are interested. First attempts at computational linguistics can be described as interdisciplinary and empiricist-centered. Noam Choamsky's generativist approach brought about a breakthrough in the field, drifting the nature of its interests. Since this breakthrough around 1960 and then on until 1985, computational linguistics was indeed mostly interested in creating "rule-based" models to describe all the linguistic phenomena surrounding us. (Jurafksy and Martin 2006) However, closer to the end of the century, linguists' attention drifted back to "data-driven" models, supported by statistical ground knowledge. During this period, computational linguistics was driving away from its theoretical roots and the field was more and more being dominated by "technicians". However, closer to our days, merely statistical approaches have been criticized, since they need too much data to be effective. Consequently, more recently, hybrid models implementing both "rule-based" and statistical algorithms have been developed. (Tamburini 2008)

1.1 Computational linguistics - Different approaches

1.1.1 Rationalist approach

A rationalist approach to the study of language is characterized by the belief that a great part of knowledge in the human mind is not derived by senses, but it is actually innate to every individual. Noam Chomsky used to be the main proponent of a rationalist approach to language; he theorized a generative grammar, which is a set of detailed rules that fully describes the innate and detailed knowledge of language every one of us possesses at birth. Chomsky's theory makes a crucial distinction between *linguistic competence* and *linguistic performance*, the first representing the set of rules previously described that we all unconsciously know, and the second representing the actual manifestation of those rules, the real language production. The aim of rationalist approaches is describing *linguistic competence*(Manning and Schütze 1999)

1.1.2 Empiricist approach

. An empiricist approach also postulates some degree of innateness of language in the human brain. There must be foundations on which to build knowledge, one cannot start from a clean slate. However, empiricists believe detailed knowledge is progressively acquired through life and is derived by more basic bits of knowledge, starting from the senses. The main difference underlying these two approaches is the object of their interest. While the first describes *linguistic competence*, the second aims at describing *linguistic performance*. Hence, an empiricist approach will draw a specific model of language starting from a general model which will be gradually modified by a large amount of input data. This data is generally organized into corpora(ibid.).

Corpora

A corpus is a special collection of textual, speech or multi-medial material collected according to a set of criteria. Generally, a corpus should be representative of the

population it aims to describe, besides it should serve the scope for which it was crafted (McEnery and Wilson 2001). In order to create a representative corpus of teenager spoken language in 2010's Italy, it would be misleading to add texts from Dante's *Inferno*, since it would not be representative at all of the way teenagers speak nowadays.

Moreover, a corpus should be balanced enough to include all the different types of text it aims to describe in an amount which is proportional to the total of texts.

Most importantly, a corpus should be able to answer the questions we are trying to answer, and should be representative of the population that we are studying (LancasterUniversity n.d.).

Limitations - Zipf's Empiricist Law

In his book *Human Behavior and the Principle of Least Effort*, Zipf argues that there is a unifying principle of humanity: we all act so as to minimize our average rate of work. This principle has consequences in every human behavior. There is evidence for it in Zipf's empirical law, which states that given an authentic corpus, the frequency of each word in it is inversely proportional to the position which the given word occupies in a rank listing all the words of the text in descending order, according to their frequency in the given corpus. As a consequence, lexical words, those which convey the meaning of a text, are way fewer than grammatical words, which do not say much about the text. Thus, statistical approaches have to face the problem of scarcity of resources from which to extract information.(Manning and Schütze 1999)

1.2 Grammaticality

1.2.1 Is grammaticality really enough?

Linguistics is interested in what people say, the way they do it, and the context in which they do it. Grammaticality has to do with the well-formedness of a sentence. A prescriptive grammar aims at distinguishing grammatical sentences from

ungrammatical ones, according to a set of rules which is thought to underlie the language competence of each speaker. However, grammaticality can be a deceptive concept, because even sentences like "Colorless green ideas sleep furiously" can be considered, in principle, grammatical. (Chomsky 1957)

Moreover, judging grammaticality becomes increasingly difficult as we proceed in our investigation up to the context level. Besides, grammaticality does not really say much about the way people speak. Firstly, because native speakers generally speak in a grammatical, correct way; conversely, conventionality is much more of a "rule" when it comes to real speaking: people use certain words and expressions simply because they hear those words and expressions more frequently. Secondly, language is subjected to constant modification and, therefore, a set of fixed rules would not be responsive enough to an ever-changing language.

1.2.2 Do we even need grammaticality at all?

According to the previous paragraph, grammaticality looks like a useless rule which theorists of language designed to their own pleasure.

However, grammar is actually a very useful instrument. As stated before, every speaker speaks in a sort of grammatical, correct way, because they unconsciously know which rules they need in order to make up a sentence and they are able to judge whether an utterance is ill-formed or not. Grammar comes in handy because it provides a framework of the language, clarifying what those rules and hidden structures we all unconsciously know are. For instance, we are all able to sense that the sentence "The worst part and clumsy looking for whoever heard light" is ungrammatical, but maybe not everybody is able to explain why.

> Here's an informal (and simplified) statement of how coordination works syntactically: Coordinate Structure: if *v1* and *v2* are both phrases of grammatical category *X*, then *v1 and v2* is also a phrase of category *X* (Bird, Klein, and Loper 2009).
>
> Briefly, we cannot conjoin a noun phrase and an adjective phrase.
> This is just a simple example, however, drawing a set of rules and recognizing

the existence of a grammar can be useful in a series of operations and fields. For instance, we could not be able to study phonetics and identify phonetic problems in children if we did not know what a phoneme is, namely if phonology did not existed. We need to categorize language in order to study it, in order to make clear which are the problems we want to study. But, at the same time, we need statistical models to study big amounts of data, in order to understand what people *usually* say, to understand what are the *conventionalities* of a given language. This has not a lot to do with the approach we want to adopt, whether it be rationalist or empiricist. Rather, it has to do with the very nature of our matter of study: language.

Chapter 2

The ambiguity of language and dependency parsing

2.1 The ambiguity of language

Language, or rather, linguistic data, is a strange entity which undergoes constant changes. Let's try to make a thought experiment and imagine we have a gigantic corpus which includes all the sentences ever uttered or written in Italian, from 1945 up to now. We can try and gather all the sentences that have been produced ever since 1945, yet our corpus will not include all the sentences that could be, or could have been, produced. This is due to the productivity of language: we can easily make up sentences that had never been thought of before, constantly creating new meanings (Manning and Schütze 1999).

As a consequence, we cannot draw general rules of a language if we admit there is no finite language. New combinations may arise and our rule would not be descriptive enough of that language we decided to circumscribe. Generativists may argue that we would just have to create a model which is descriptive enough to react to changes in language. However, even though we assumed that a language is a finite, circumscribable set of sentences, ambiguity of interpretation would still be a problem (Bird, Klein, and Loper 2009). For instance, the sentence "The dog saw a man in the park" could be interpreted in two ways, as shown in figure 2.1. In figure 2.1(a) the seeing action happens in the park, while in figure 2.1(b) the man

is in the park and the agent of seeing is somewhere outside of it.

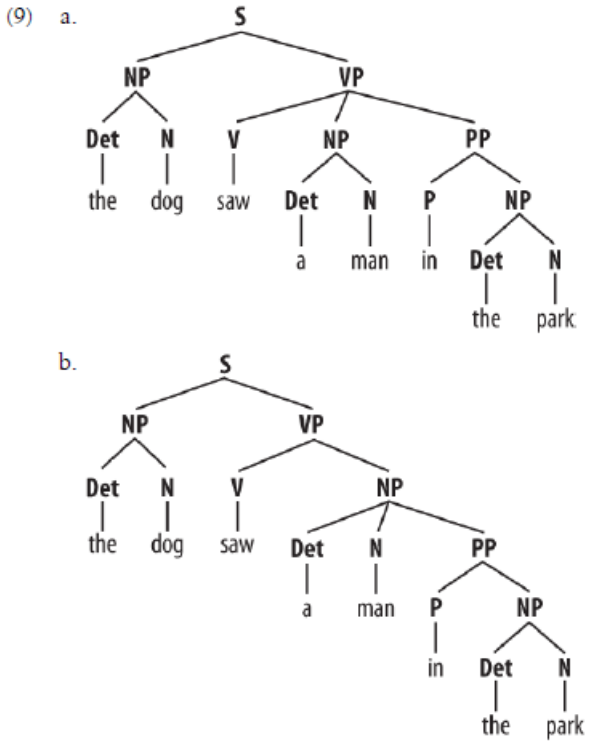


Figure 2.1: Sentence structure ambiguity.

The objective of a linguistic model is that of describing as many linguistic productions of a given language as possible, thus being able to interpret as many sentences as possible. And of course, such a model needs a starting point, a reference grammar, a set of rules of the given language. It is not possible to build an interpretation scheme starting from a *tabula rasa*. However, evaluating the ability of interpretation of such a model goes beyond the scope of this thesis (Bird, Klein, and Loper 2009).

2.2 Parsing: dependency parsing

As written above, a set of grammatical rules is needed in order to interpret a sentence. Particularly, closing in on the objective of this thesis, this set of rules is needed to interpret the possibly ambiguous structure of a sentence. We call "parser" one of the many ways a sentence structure can be interpreted starting from a reference

grammar. There are two main ways of parsing: constituency parsing and dependency parsing.

In constituency parsing, the syntactical relations of a sentence are specified, in a top-down manner, by means of constituent structures (or trees). An example is shown above in figure 2.1.

On the other hand, in dependency parsing relations among words are shown through labeled arrows which go from the **head** of the given construction, i.e. of the given relation, to the **dependent** of the same relation. The starting point of the parsing, the root, is the main verb, which serves as the head of the whole sentence. Its dependent will be a word which is most closely related to it. This word in turn will be the head of another dependent. The bottom-end dependents are usually grammatical words, such as "of", "the", "with", "my", and so on. In figure 2.2 an example of a dependency tree is shown. Generally, dependency parsing results to be more effective in processing linguistic data compared to constituency parsing, especially speech data (ibid.).

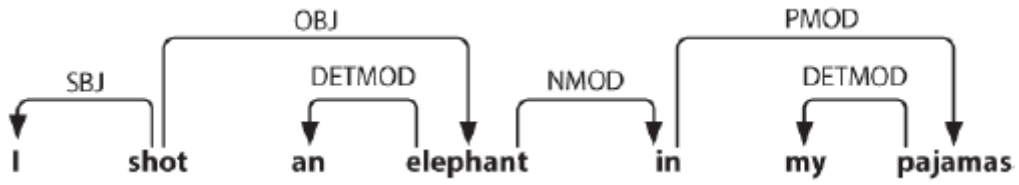


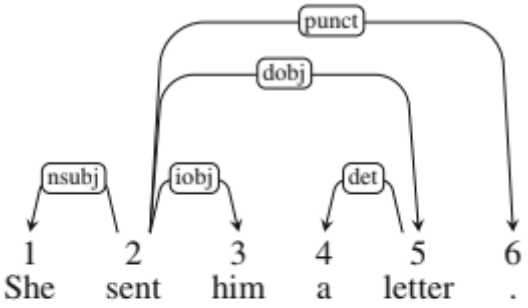
Figure 2.2: I shot an elephant in my pajamas.

Here follows a list of criteria which helps determining which is the head H and which is the dependent D of a construction C (ibid.).

1. H determines the distribution class of C; or alternatively, the external syntactic properties of C are due to H.
2. H determines the semantic type of C.
3. H is obligatory while D may be optional.
4. H selects D and determines whether it is obligatory or optional.
5. The morphological form of D is determined by H (e.g., agreement or case government.)

2.2.1 De Facto Standards: CoNLL and Dependency Annotation

A dependency tree is basically made of a series of words, to each of which a syntactic head and a dependency annotation are assigned. Most parser and treebank formats use a representation in which each word of a sentence is assigned with a head index and a dependency label. Treebanks are collections of correctly parsed sentences which use specif part-of-speech tags. They are generally used to build up parsers. An example of format is the Malt-TAB format, which represents a word token, giving to it four attributes: index, word form, head index, dependency label (fig. 2.3).




```
1 She    2 nsubj
2 sent   0 root
3 him    2 iobj
4 a      5 det
5 letter 2 dobj
6 .      2 punct
```

Figure 2.3: Malt-TAB format.

An important step in the development of shared tasks in dependency parsing was the CoNLL-X format, which was devised using data sets from 13 languages. CoNLL-X format had to be expressive enough to summarize the annotations of the 13 native annotation formats. This is the reason why it represents a word token using 13 different attributes. An example is shown in figure 2.4.

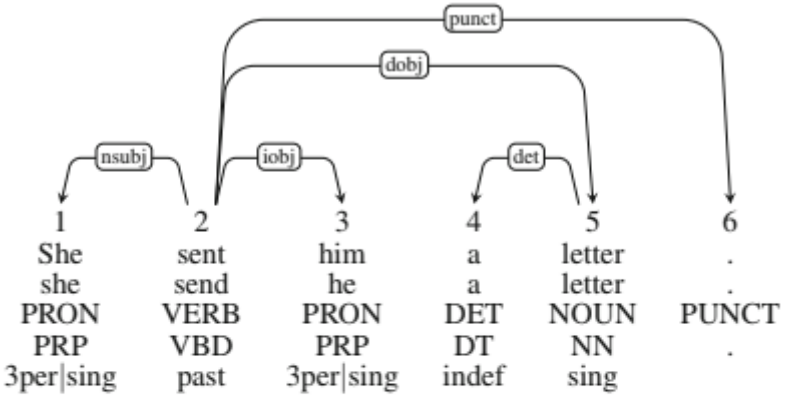
1. ID: Token counter, starting at 1 for each new sentence.
2. FORM: Word form or punctuation symbol.

3. LEMMA: Lemma or stem (depending on the particular treebank) of word form, or an underscore if not available.
4. CPOSTAG: Coarse-grained part-of-speech tag, where the tagset depends on the treebank.
5. POSTAG: Fine-grained part-of-speech tag, where the tagset depends on the treebank. It is identical to the CPOSTAG value if no POSTAG is available from the original treebank.
6. FEATS: Unordered set of syntactic and/or morphological features (depending on the particular treebank), or an underscore if not available. Set members are separated by a vertical bar (|).
7. HEAD: Head of the current token, which is either a value of ID, or zero (0) if the token links to the virtual root node of the sentence. Note that depending on the original treebank annotation, there may be multiple tokens with a HEAD value of zero.
8. DEPREL: Dependency relation to the HEAD. The set of dependency relations depends on the particular treebank. The dependency relation of a token with HEAD=0 may be meaningful or simply ROOT (also depending on the treebank).
9. PHEAD: Projective head of current token, which is either a value of ID or zero (0), or an underscore if not available. The dependency structure resulting from the PHEAD column is guaranteed to be projective (but is not available for all data sets), whereas the structure resulting from the HEAD column will be non-projective for some sentences of some languages (but is always available).
10. PDEPREL: Dependency relation to the PHEAD, or an underscore if not available. (Ide and Pustejovsky 2017)

Projective parsers

A dependency tree is considered to be *projective* if, when all the words of a sentence are put in linear order, preceded by *root*, edges above words can be drawn without

crossings.



1	She	she	PRON	PRP	3per|sing	2	nsubj	-- --
2	sent	send	VERB	VBD	past	0	root	-- --
3	him	he	PRON	PRP	3per|sing	2	iobj	-- --
4	a	a	DET	DT	indef	5	det	-- --
5	letter	letter	NOUN	NN	sing	2	dobj	-- --
6	.	.	PUNCT	.	--	2	punct	-- --

Figure 2.4: CoNLL-X format.

2.2.2 Dependency parsers: LinguA - Linguistic Annotation Pipeline

There are several dependency parsing tools which are trained on the Italian language, two examples of which are Tule, the Turin University Linguistic Environment, which includes a chunk-rule based dependency parser (http://www.tule.di.unito.it/ n.d.), and Tint, a dependency parser based on Stanford CoreNLP (Fondazione-BrunoKessler n.d.). However, I am now going to more accurately describe another dependency parser: LinguA, which I used during reasearch around the thesis' case study.

LinguA is a linguistic annotation pipeline which was devised and developed by the ItalianNLP Lab at the Computational Linguistics Institute "Antonio Zampolli" (ILC-CNR). LinguA combines rule-based and machine learning algorithms. The annotation process is divided into four steps:

1. Sentence splitting: sentences are split according to punctuation. Full stops, exclamation points, question marks and also carriage returns will be processed as sentence splitting points.

2. Tokenization: process of dividing up a text into units (tokens), which can be words, numbers or punctuation marks. (Manning and Schütze 1999)

3. Part-of-speech tagging and lemmatization: PoS tagging is the process of labeling each word of a sentence with its appropriate part of speech (noun, verb, preposition...), according to different degrees of specificity; lemmatization is the process of labeling a word with the related lemma, which is the non-inflected dictionary entry from which every word "stems".

4. Dependency parsing: this process was described in detail above in the chapter, section 2.2. (ItaliaNLP-Lab n.d.)

LinguA analyses the inserted text and provides an analysis visualization in CoNLL-X format, where:

- sentences are separated by a blank line;
- each token starts on a new line and it is annotated with the following linguistic information: lemma, coarse and fine grained part-of-speech, morphological features, syntactic dependency information. (ibid.)

For Italian, LinguA uses the ISST-TANL mopho-syntactic and dependency tagset, while for English it uses the standard Penn TreeBank tagset. Example figures from LinguA will be shown later on in the thesis, where real text samples will be provided.

Chapter 3

Developmental Language Disorder

A definition of DLD

Developmental Language Disorder is a clinical condition for which children can present problems on spoken language ability which are severe and persistent enough to cause long-term complications in daily life communication and education attainment (Association 2013).

A child suffering from DLD do not present any differentiating conditions (section 3.1.3): they hear normally and show no neurological damage or disease. Plus, they score at age-appropriate levels on tests of nonverbal intelligence. Moreover, a child suffering from DLD presents areas of weakness in language which can be associated to all of those language abilities which pose difficult challenges also for a child with normal development (Leonard 2014). This has led researchers to think of DLD as the bottom end of a language aptitude continuum, rather than as the "ill half" in a disorder-nondisorder dichotomy. An additional proof for this interpretation resides in the fact that

> (...) heritability estimates for language test scores and nonword repetition scores in particular appear to be no different for the lower end of ability (where children with SLI reside) than for the average range. (ibid.)

Lastly, in the language domain, DLD cannot be divided up into neatly categorized subtypes. Weaknesses in various areas of language domain may overlap.

Besides, DLD is language-dependent. Hence, problems may differ according to the particular language the child is acquiring.

From here on a more detailed description of DLD will be provided. (Bishop 2017)

The problem with terminology

Even though I wrote *Developmental Language Disorder*, the correct terminology to be used has not always been the same. As a matter of fact, in medical literature, DLD has been referred to also as SLI (Specific Language Impairment). However, according to a 2017 DELPHI Consensus Study on English terminology, the preferred term is now DLD. According to this Consensus Study, the term SLI had become controversial because it seemed not to reflect clinical realities and excluded many children from the study. The word "specific" has been criticized as particularly misleading, because DLD can co-occur with conditions related to other psycho-physical abilities, such as attention or non-verbal communication. A similar distinction of preference of use has been made in Italian between *Disturbo Specifico del Linguaggio*, literally translated as "Specific Language Disorder", and *Disturbo Primario del Linguaggio*, which is "Primary Language Disorder". "Developmental" here refers to the fact that this condition emerges during the child's development, rather then being derived by another specific bio-medical condition (ibid.).

3.1 Digging up details

3.1.1 Late talkers

Late talkers are children who present an underdeveloped use of language compared to children of the same age. As pointed out in section 3.1.2, it is difficult to predict whether a child will develop serious conditions related to language disorder. However, this clinical study relies on the fact that, despite the difficulty of finding predictive values of language disorder, the earlier a child starts a therapy, the higher the probability he or she will catch up and develop a typical level of language

competence, because of higher brain plasticity.

Late talkers usually present the following characteristics:

- Underdeveloped use of language;
- Slow growth of expressive vocabulary, which is slow growth in the ability to express one's emotional state through specific linguistic elements, such as interjections or hyperbolas;
- Underdeveloped morpho-syntactic level at 30 months of age;
- Verbal comprehension deficit;
- Lower intelligibility compared to peers.

3.1.2 DLD and age

Prognostic indicators may vary according to the age of the child. Approximately, the younger the child, the more difficult it is to predict whether he or she will develop long-term language disorders.

Under three years of age, it is particularly hard to predict whether a "late talker" will develop serious conditions, since many toddlers who presents linguistic underdevelopment before 3 years of age eventually catch up (Chilosi 2019). However, this is the age under which therapy can have the most satisfactory results, because of high brain plasticity. Risk factors of developing DLD include a positive family history of language or literacy problems, and preterm birth.

After 3 years of age and up to 4, prediction accuracy increases. The greater the number of linguistic areas impaired, the higher the probability that the language disorder will persist in school age. When problems are still present at 5 years of age, they are likely to persist. Children who start school with language disorders are at risk of literacy problems and poor academic results (Bishop and Edmunson 1987).

3.1.3 Differentiating conditions

Differentiating conditions are biomedical conditions in which language disorder may occur as a consequence of a complex condition which is not strictly related to lan-

guage, but which can hinder it.

> Differentiating conditions include deafness, brain injury, acquired epileptic aphasia in childhood, certain neurodegenerative conditions, cerebral palsy and oral language limitations associated with sensori-neural hearing loss, as well as genetic conditions such as Down syndrome. We also include here children with autism spectrum disorder (ASD) and/or intellectual disability because these conditions are commonly linked to genetic or neurological causes (Bishop 2017).

In these cases, the correct terminology to be adopted is "Language disorder associated with X". When such condition is diagnosed, treatment will be different than when DLD is not associated with any differentiating conditions, and will have to take into account the specificity of each case (ibid.).

3.1.4 Co-occuring disorders

Co-occuring disorders are impairment in cognitive, sensori-motor and behavioral domains which can co-occur with DLD, and whose link with language disorder is yet unclear. These include disorders related to attention, such as ADHD, motor problems, reading and spelling problems, developmental dyslexia, and others.

According to clinical research, most children actually present a mixture of problems, rather than a set of distinct conditions. For this reason, different experts may label the same condition in different ways (ibid.).

3.1.5 Areas of language and DLD

Developmental Language Disorder is a heterogeneous category which encompasses a series of impairments in different areas of language domain. There is no common agreement in terminology regarding the various subtypes of DLD. A list of the areas of language domain which can be impaired will follow. It is important to point out, however, that there is no clear distinction between one subtype and another, problems in different areas may indeed overlap (ibid.).

Phonology

Phonology is the branch of linguistics which studies speech sounds and ways to categorize them. Phonetic impairments may pose serious difficulties in communication and, as a consequence, in child's education.

One of the most common characteristic of phonetic impairment is the inability of distinguishing minimal pairs, which are pairs of words which are mutually distinctive because of one and only one phoneme, which indeed contrast the meaning of the words in the pair. Such words are, for instance, *pin* (/pɪn/) and *bin* (/bɪn/) in English, or *tana* (/'tana/) and *lana* (/'lana/) in Italian. This type of impairment hinders comprehension and reproduction of meaning (ibid.).

Morpho-syntax

Morpho-syntax deals with the categorization of variation in shape and structure of words and sentences. Variation in the morphology of words, or rather, inflection, can convey meanings such as gender, number, or case of a given word. Syntax variations have to do with the meaning of a sentence, and its discourse role: questions in constrast with affirmation, for example.

Children who present morpho-syntax impairment, have difficulties in understanding the meaning of sentences which are marked by grammatical contrast. For instance, distinguishing a sentence in the past tense from one in the present tense, or distinguishing grammatical from ungrammatical sentences. (ibid.)

Semantics

Semantics is the branch of linguistics which aims at categorizing words according to their meaning and to the their paradigmatic positioning, which is the way they relate to other words outside of the sentence they are found in.

Children with impairment in the semantic area may present difficulties in understanding and organizing the meaning of words. A task in which they present problems performing is "word finding": despite having some knowledge of the meaning of some words, they struggle to produce those words (ibid.).

Pragmatics

In linguistics, pragmatics aims at describing and categorizing the principles in language which determine literal and non-literal meanings of sentences according to their context of use. An important ability which underlies pragmatics is Speech Acts, which is performing an action through speech. For example the sentence "I now pronounce you man and wife" declares that a man and a woman officially became husband and wife (Austin 1955).

Children with pragmatic impairment present difficulties in linking production and comprehension of language to the context they find themselves in. For example, they can be too literal in their expressions, or giving too much information, even when it is unnecessary because of the context. Or else, they may not be able to understand social cues and adapt their comprehension or language production to them. Other aspects are related to language prosody, which can result artificial (Bishop 2017).

Discourse

In linguistics, discourse is related to the fact that the meaning of a text is derived by the understanding of a complex structure made of many different bits (sentences).

Children suffering from DLD may lack the ability to form in their mind the meaning of a coherent whole starting from a sequence of fragmented sentences. This may result in the inability of understanding a text and also producing it (ibid.).

Learning and memory

Research literature has shown that children with language disorder may struggle in memorizing sequences of words or sounds in a short delay and in reproducing them. One specific problem is represented by their inability of reproducing non-words (nonsense words) of three to four syllables (ibid.).

3.1.6 Cross-linguistic differences

According to Leonard (2014), children suffering from DLD present different weaknesses according to the language they are exposed to. This is due to the specific characteristic of each language. Leonard adduced a series of examples to his theory.

English, for example, is a poorly inflected language. For instance, there is almost no case marking, except for pronouns, gender in nouns and adjective is also not marked, verbs have few forms: past-tense-marking (irregular and regular declensions) and third person 's'-marking. As a consequence of this inherent characteristic of English, many English-speaking children suffering from DLD present (and it has been well-documented) deficits in using tenses and agreement inflections.

However, in Romance languages such as Italian and Spanish, which are highly inflected languages and whose inflection system is quite simple and phonologically clear, children rarely have difficulties in using tenses and agreement inflections. Nevertheless, they present other kind of problems which reflect the specific characteristics of Romance languages. For instance, Italian children with DLD struggle with the use of unstressed direct object pronouns which, in Italian, must precede the verb rather than follow it. (e.g., *Mario compra il pane* [*Mario buys bread*], but *Mario lo compra* [Mario it buys]).

Another example is the verb-second property of languages, such as German, Dutch and Swedish. The normal wording is *subject+verb+object*. However, if the first word of a sentence is a word other than the subject, the verb must always be the second element (e.g., *Birgitta äter glass [Birgitta eats ice cream]*, but *Sen äter Birgitta glass [Then eats Birgitta ice cream]*).

A further evidence for cross-linguistic difference is the fact that bilingual children present almost the same characteristics as monolingual children. Hence, the ambient language influences the kind of problems a child will present in the given language. (ibid.)

Chapter 4

Case study - Back to Italy

This last chapter of the thesis will examine a clinical study that was conducted in Italy, more specifically in a local healthcare center in Tuscany (USL Toscana Centro). More precisely, it will look at how the data from this clinical study were analyzed and elaborated in order to craft this thesis. I will briefly comment on the results that were drawn out of the study itself, and then I will describe the learning process I went through while studying and working on the data provided by this clinical study.

My comment will be limited to the statistical relevance of the data and subsequent processing. While writing the thesis, not only could I acquire new competences, primarily in the field of computational linguistics and also around a topic which is more related to neuroscience and speech therapy, but I also managed to create some useful instruments for oral text elaboration in text therapy. Firstly, I created a small corpus of dependency-parsing-annotated speech-therapy-relevant texts. Secondly, I demonstrated the benefit of using a type/token ratio calculator in speech therapy.

4.1 Case study: a clinical study on late talkers

4.1.1 Description of the study

The clinical study to be discussed was conducted in Italy on a group of children from 24 to 36 months of age. The group included 18 children (10 male and 8 female), evenly divided into typically developing children and late talkers. The first half served as a control group (Ferrari, Gagliardi, and Innocenti 2018).

According to the premises of the clinical study, particular attention should be brought to children' gestures, since a high number of gestures represent a predicting factor of recovery.

4.1.2 Methods

The clinical study focused on analyzing the communicative interactions between a care-giver (a parent) and their child. It was a semi-structured interaction conducted in an ecological context. An ecological context, in research methodology, is an environment which aims at recreating a real-life situation in order to increase the degree of validity of the results to be obtained in a given study. The interaction was thus divided into three phases (Ferrari 2018):

- Functional-symbolic game, which involves the reproduction of a series of actions carried out by the adult;
- Construction game: particularly useful for those children who rely on practical activities to develop language. It involves the use of "construction" abilities with the help of the care-giver;
- Book reading: the activity involves the reading of two simple books with the help of the care-giver.

4.1.3 Clinical study results

The main differences between the control group and late talkers are related to turn taking and gestures, lexical composition and morpho-syntactic predictive indicators

of a language disorder, and care-giver communicative style.

In the late talkers group, the care-giver usually intervenes more often than in the control group, in order to hint the child or to fulfill moments of silence. Gestures did not show statistical relevance, yet they can be a reliable predicting factor, as pointed out above in the chapter.

Late talkers and tipically developing children present differences in lexical composition, the mopho-syntactic elements of which can be predictive of a language disorder. For instance, late talkers tend to use many more holophrasis, which are single-word sentences used to express meaningful thought (e.g., Italian child-word "pappa" [literally: food] used to express: "I want to eat".) They also use many more sounds and tend to use less complex word combinations.

In the control group, care-givers usually use a "tutorial" communicative style, they encourage the child's language production. In the late-talker group, instead, parents tend to discourage the child (ibid.).

4.2 Computer programs

In the "preparation" phase of the thesis, a series of programs has been used. The main purpose of their usage was processing text. The starting point of the elaboration was represented by transcriptions from the speech therapy sessions which were part of the clinical study described in the previous section. While the result of the processing is an annotated text, which can be used for further calculations and analysis. A more detailed description of the programs and the kind of semi-automatic text annotation that was carried out will follow.

4.2.1 L-AcT format and ELAN

All the sessions of therapy were recorded and some parts of each were transcribed according to the L-AcT format, using the software ELAN.

L-AcT format is a version of the standardized CHAT format, enriched with tagging of prosodic parsing. This means that all the prosodic breaks, namely the boundaries of an utterance, are signaled through a series of pre-established symbols,

as shown below in figure 4.1 (Cresti and Moneglia 2018).

L-AcT diacritics for the annotation of prosodic structure	
Prosodic break perceptively relevant prosodic variation in the speech continuum such as to cause the parsing of the flow into discrete prosodic units	
Terminal Perceptual criterion: a competent speaker assigns to it the quality of concluding the sequence	// ? (with interrogative prosodic profile)
Non terminal Perceptual criterion: a competent speaker assigns to it the quality of being non-conclusive	/
False Start/retracting with repetition Non terminal prosodic break caused by a false start or retracting	[/]
Unintentionally interrupted sequences The speaker's program is broken; the interpretability of the sequence can be compromised	+
Empty pause Temporary silent hesitation or stop in the speech flow (lower-bound threshold: 250 ms)	#

Figure 4.1: Symbols in L-AcT format

ELAN is a computer tool which can be used, as pointed out above, to transcribe from video or audio resources. For the purpose of this thesis, the texts that had been transcribed were subsequently imported (extracted) from ELAN as .txt files, and then opened in Visual Studio Code, which is, as the name quite explains for itself, a text and code editor. In figure 4.2 an example of how the text looks like in Visual Studio Code, right after the import from ELAN.

```
1   cosa si costruisce?
2
3   # vai//
4
5   tieni / inizia//
6
7   # si fa una torre?
8
9   si fa una torre / Chiara?
10
11  e che cosa vuoi fare?
12
13  la torre//
14
15  così / guarda//
```

Figure 4.2: Text imported from ELAN opened in Visual Studio Code

4.2.2 Python and text cleaning

Python is an object-oriented, high-level programming language. This means that, in Python, apparently incompatible units of code can communicate with one another, and that such a programming language is pretty easy to use, because its logic is much closer to human logic than to computer logic. Hence, the objects it uses are quite intelligible (PythonSoftwareFoundation n.d.)(Rossum 1995).

The main purpose Python served in the preparatory work to this thesis was cleaning up the text from all the L-AcT "markings", in order to prepare it for further elaboration. Figure 4.3 shows a part of the Python script which converts L-AcT marks into either traditional punctuation marks, or blank spaces.

```
a =text.replace("//", ".") # substitutes LABLITA // with . for parsing
b =a.replace("/", ",") # substitutes LABLITA / with , for parsing
c =b.replace("#", "") # removes hash marks
d =c.replace("[", "") #substitutes LABLITA [ with a blank
e =d.replace("]", "") # substitutes LABLITA ] with a blank
f =e.replace("xxx", "") # removes unintelligible parts
g =f.replace("hhh", "") # removes paraverbal parts' annotations
h =re.sub('[A-Z0-9]{3,}', '', g) # removes speakers' names
i =re.sub(r'[\n]{2,}', r'\n', h) # removes new line blanks
```

Figure 4.3: Python example from the script that was used.

4.2.3 LinguA

At the end, the text was extracted from ELAN into a .txt file. Then the Python script was run. As a consequence, all the texts imported from ELAN input the Python script as "raw text" and output it as clean text. This clean text appears as we would normally see dialogues on a book. Thus, it is now ready to be processed by LinguA, the linguistic annotation pipeline devised and developed by ItalianNLP lab which has been already described in subsection 2.2.2.

LinguA output an annotation of the input text which looks as shown in figures 4.4 and 4.5.

	ID	Token	Lemma	C-POS	F-POS	Morphosyntactic feats	HEAD	DEP
1	1	cosa	cosa	P	PQ	num:s gen:n	3	subj
	2	si	si	P	PC	num:n gen:n per:3	3	clit
	3	costruisce	costruire	V	V	num:s mod:i per:3 ten:p	0	ROOT
	4	?	?	F	FS		3	punc
2	1	#	#	S	SW	num:n gen:n	2	subj
	2	vai	andare	V	V	num:s mod:i per:2 ten:p	0	ROOT
	3	/	/	X	X		2	punc
	4	/	/	X	X		0	ROOT

Figure 4.4: Annotated text as shown online in LinguA.

```
1   1   la   il   R   RD   num=s|gen=f 2   det
2   2   nuvola   nuvola   S   S   num=s|gen=f 10   subj
3   3   Olga   Olga   S   SP   _   2   mod
4   4   ,   ,   F   FF   _   2   con
5   5   e   e   C   CC   _   2   con
6   6   l'   il   R   RD   num=s|gen=n 7   det
7   7   uccellino   uccellino   S   S   num=s|gen=m 2   conj
8   8   Ugo Ugo S   SP   _   7   mod
9   9   ,   ,   F   FF   _   2   punc
10   10   vanno   andare   V   V   num=p|per=3|mod=i|ten=p 0   ROOT
11   11   a   a   E   E   _   10   arg
12   12   vedere   vedere   V   V   mod=f   11   prep
13   13   il   il   R   RD   num=s|gen=m 14   det
14   14   mare   mare   S   S   num=s|gen=m 12   obj
15   15   .   .   F   FS   _   10   punc
```

Figure 4.5: Annotated text from LinguA as shown in Visual Studio Code.

4.2.4 Python and statistical elaboration: type/token ratio

However, Python served also another scope: calculating the type/token ratio of annotated texts. The aim of this calculation is comparing the richness of vocabulary in the verbal productions of the groups of children and parents.

The type/token ratio represents the total number of unique words (types) divided by the total number of words (tokens). This ratio was automatically calculated using a python script, thus eliminating the need to manually calculate it. The results are shown in the table below, which includes TTR of late talker children and their parents, and TTR of normally developing children and their parents.

	TTR lt children	TTR lt parents	TTR nd children	TTR nd parents
	0.2909	0.2898	0.2656	0.2139
	1.0	0.2358	0.29965	0.2158
	0.1979	0.2210	0.2571	0.2517
	0.2868	0.2195	0.2489	0.2923
	0.3047	0.1837	0.2992	0.2238
	0.0929	0.2260	0.2016	0.2283
	0.2949	0.1685	0.2764	0.2646
	0.3732	0.1844	0.3413	0.2068
	0.1292	0.2047	0.3713	0.2492
mean	0.3301	0.2179	0.2846	0.2385
SD	0.2669	0.0353	0.0507	0.0281

The table also shows the mean and the standard variation of the data from the four groups. P-value, which indicates the degree of representativeness of the data compared to reality (its likelihood to correspond to real data), is 0.2942 for the late talker child-parent group, and 0.7301 for the other group.The higher the p-value, the lower the representativeness of the data.

Although there is not enough evidence to build up a statistically meaningful model, because p-value is over 0.05, it is interesting to note that standard variation is much higher in the late talkers, while the normally developing group of children is more homogeneous.

The whole process of semi-automatically text annotation combined with the type/token ratio speeds up the work of the speech therapist, who usually has to manually calculate the ratio.

Conclusion

In this thesis I examined the topic of computational linguistics, introducing the main ideas and the problems which this science went through and those which it is still facing today. I then focused on dependency parsing, explaining its main characteristics.

In the third chapter, the analysis moved on to Developmental Language Disorder. I introduced the topic and described the characteristics of DLD and the main problems researchers on the topic have been facing.

In my last chapter I described how I worked in order to get to the final result: this very thesis. Besides, I showed how computational linguistics and speech therapy may be linked, in doing so I showed how to craft two tools which could turn out to be useful in the work of the speech therapist: a small annotated corpus containing oral texts from speech therapy sessions in which children suffering from DLD interact with one of their parents, and a python script which calculates the type/token ratio of texts.

To conclude, I presented the topic of computational linguistics and demonstrated how field-related competences and applications may be used in fields which are not directly related to automatic language processing, such as speech therapy. The result is an improved experience in research and diagnosis.

Bibliography

Aprosio, Alessio Palmero and Giovanni Moretti (2016). *Italy goes to Stanford: a collection of CoreNLP modules for Italian.*

Association, APA - American Psychiatric (2013). *Diagnostic and statistical manual of mental disorders*. Ed. by American Psychiatric Publishing. Fifth. Washington DC - London.

Attardi, Giuseppe and Felice Dell'Orletta (2009). "Reverse Revision and Linear Tree Combination for Dependency Parsing". In: *NAACL-HLT 2009 - North American Chapter of the Association for Computational Linguistics - Human Language Technologies*. Boulder, Colorado, pp. 261–264.

Attardi, Giuseppe, Felice Dell'Orletta, et al. (2009). "Accurate Dependency Parsing with a Stacked Multilayer Perceptron". In: *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian*. Reggio Emilia, Italy.

Austin, John Langshaw (1955). *How To Do Things With Words*. Ed. by Oxford University Press.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. Ed. by O'REILLY. Sebastopol, California.

Bishop, Dorothy V. M. (2017). "Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology". In: *Journal of Child Psychology and Psychiatry*. Ed. by John Wiley and Sons Ltd.

Bishop, Dorothy V. M. and A. Edmunson (1987). "Language-impaired 4-year-olds: distinguishing transient from persistent impairment". In: *Journal of Speech and Hearing Disorders* 52. Ed. by John Wiley and Sons Ltd, pp. 156–173.

Chilosi, Anna Maria (2019). "Which linguistic measures distinguish transient from persistent language problems in Late Talkers from 2 to 4 years? A study on Italian speaking children". In: *Elsevier*.

Chomsky, Noam (1957). *Syntactic Structures*. Ed. by Walter de Gruyter GmbH and Co. KG. Berlin, Germany.

Cresti, E. and M. Moneglia (2018). "The illocutionary basis of information structure". In: Adamou, E., K. Haude, and M. Vanhove. *Information Structure in Lesser-described Languages. Studies in prosody and syntax*. Ed. by John Benjamins. Amsterdam-Philadelphia. Chap. 13, pp. 360–402.

Dell'Orletta, Felice (2009). "Ensemble system for Part-of-Speech tagging". In: *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian*. Reggio Emilia, Italy.

Ferrari, E. (2018). *Studio delle abilità pragmatiche del bambino parlatore tardivo: analisi degli atti linguistici durante l'interazione con il genitore.* Corso di Laurea in Logopedia (abilitante alla professione sanitaria di Logopedista) – Università degli Studi di Firenze.

Ferrari, E., G. Gagliardi, and M. Innocenti (2018). *Studio delle abilità pragmatiche del bambino parlatore tardivo: analisi degli atti linguistici durante l'interazione con il genitore.* Giornate di studi scientifici sul linguaggio, Rovereto.

FondazioneBrunoKessler (n.d.). *Tint Dependency Parsing*. URL: `http://tint.fbk.eu/parsing.html`.

http://www.tule.di.unito.it/ (n.d.). *TULE - Turin University Linguistic Environment*. URL: `http://www.tule.di.unito.it/`.

Ide, Nancy and James Pustejovsky (2017). *Handbook of Linguistic Annotation*. Ed. by Springer. Dordrecht, The Netherlands.

ItaliaNLP-Lab (n.d.). *LinguA*. URL: `http://www.italianlp.it/demo/linguistic-annotation-tool/`.

Jurafksy, Daniel and James H. Martin (2006). *Speech and Language Processing. An introduction to natural language processing, computational linguistics, and speech recognition*. Ed. by Pearson College Div.

LancasterUniversity (n.d.). *Corpus linguistics: Representativeness, balance and sampling*. URL: https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/A02.pdf.

Leonard, Laurence B. (2014). *Specific Language Impairment Across Languages*. Author Manuscript. National Institutes of Health. Department of Speech, Language, and Hearing Sciences Purdue University.

Lesmo, Leonardo (2007). "Il parser basato su regole del Gruppo NLP dell'Università di Torino". In: *Intelligenza Artificiale*. Chap. 4, pp. 46–47.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Ed. by Massachusetts Institute of Technology. United States of America.

McEnery, Tony and Andrew Wilson (2001). *Corpus Linguistics*. Ed. by Edinburgh University Press Ltd. Second. United Kingdom.

PythonSoftwareFoundation (n.d.). *What is Python? Executive Summary*. URL: https://www.python.org/doc/essays/blurb/.

Rossum, Guido van (1995). *Technical Report CS-R9526*. Ed. by Centrum voor Wiskunde en Informatica (CWI). Amsterdam.

Tamburini, Fabio (2008). "La linguistica computazionale: un crogiolo di esperienze multidisciplinari." In: ed. by Griseldaonline, pp. 1–11.