The Information Structure–Prosody Interface

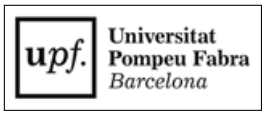
On the Role of Hierarchical Thematicity in an Empirically-grounded Model

Mónica Domínguez Bajo

TESI DOCTORAL UPF / ANY 2017

DIRECTORS DE LA TESI
Dr. Leo Wanner i Dr. Mireia Farrús

DEPARTAMENT DE TECNOLOGIAS DE LA INFORMACIÓ I LES COMUNICACIONS



To my daughter, Tesla.

Your acceptance to have in me a part-time mum, who had to often overwork at night, has made this dissertation possible.

Acknowledgements

These years from 2012 to 2017 have probably brought me the happiest and, at the same time, saddest moments in my life. I often felt as if I was rowing on an unstable little boat in the middle of the ocean under a heavy storm. But, now that the boat is finally anchored near the shore, I revisit those feelings and realize that many people were by my side in this academic and personal enterprise. I would like, most of all, to express the deepest appreciation to my thesis supervisors: Leo Wanner, who conveyed a spirit of continuous improvement and enjoyment with respect to research; and Mireia Farrús, who turned out to be a close friend apart from a devoted co-supervisor. Both of them have contributed to plot the ship's course despite my (sometimes wild and mostly adventurous) outbursts of creativity. The doctoral committee deserves the most sincere thank you for accepting the non-trivial task of assessing this PhD dissertation; I hope my writing serves as a lighthouse that leads this vessel to a safe harbor. A very special gratitude goes to Igor Mel'čuk, the "invisible" shipbuilder of my research; I wish we can make ourselves visible at some point soon.

From the DTIC-UPF crew, there are so many people to thank. I had the honor to have the most precious advice and guidance of Alicia Burga throughout these years: the annotation of thematicity and my understanding of communicative structure is the result of her dedication and her vast knowledge of linguistics. The implementation of *Praat on the Web* looks as smart as it does thank to the work of Iván Latorre and Joan Codina; besides, their advice on software architecture saved me endless hours of debugging Praat scripts. Francesco Ronzano and Joan Pere Sánchez, my office mates for three years, also guided me in technical issues with machine learning, TTS intricacies and endured my "roller-coaster emotional trips" during turbulent tempests. Down in the galley, Federico Sukno, Adrià Ruiz, Oriol Martínez, Vanesa Vidal, Gerard Casamayor, and Beatriz Fisas shared annotation "nightmares", I mean, "tasks" of different natures, post-telco coffees and an overall super positive valence-arousal attitude; rowing as a team makes hard work so much bearable. Many people from the department and from European projects shared with me wonderful moments, here are just a few outstanding names: Aurelio Ruiz, Ralph Andrzejak, Horacio Saggion, Gerard Llorach, Azadeh Faridi, David Molero, Dominik Schiller, Florian Lingenfelser, Bianca Vieru, Lori Lamel, Stefanos Vrochidis, Eleni Kamateri, Jutta Mohr, Louisa Pragst and a long etcetera. It's so nice to go on deck with you, comrades. So nice, just to breath deeply and look over to the horizon.

My family and friends deserve a special mention. Distance can sometimes be an unsurmountable barrier, but they have all proved true bonds go beyond time and space. My mum, dad and brother were ready to lever the burden of multitasking at any time, so that I could concentrate on being a student without any extra responsibilities. To all those who resist stubbornly to persist in their friendship, I owe a piece of my heart: Rosa, Aida, Noe, Pedro, Dani, Maddi, Rubén, Maikol, Iván, Borja, Carol and Patu. Colleagues from the TALN (Roberto, Simon, Luis, Francesco, Juan, Alp, Miguel, Luz, among others) filled in what otherwise would have been sad lonely Friday evenings afterwork. A warm thank you to mums of Tesla's schoolmates, especially Nùria, Marta, Therese, Paola and María. We already hold a doctorate in motherhood: how to raise a toddler, work and survive to tell about it.

In the process of becoming aware that there is always the possibility to move on from ocean storms and lonely boats, Captain Martu has been a Master of Zen. You offered me a hand on the very verge of sinking. You may call it synergy, affection, teamwork, listening skills, profiling... I can just say I have no words (and so many at the same time) to thank you for the unconditional support, commitment and love we share on a daily basis. Even though water was breaking into the ship when you boarded, you stayed calm (as the good Captain that you are) and delicately mended the numerous parts that were broken. As deep as the ocean, as wide as the sea, the blossom of a fragrant rose stands in the middle of a harvested field, witness of a tower that reminds us there are other worlds than these.

Going back to prose, I would like to acknowledge all those senior researchers that have given me feedback in my posters and presentations in international conferences during these years: Hiroya Fujisaki who kindly took interest in my preliminary experiments in SP2014; Julia Hirschberg for her interest in both SP2014 and SP2016; Barbara Plank for her enthusiasm when she popped into my poster in COLING 2016; Andrew Rosenberg for his constructive criticism on the automatic prosody tagger; and Nick Campbell for a really warm and friendly welcome at my first international conference (SP2014, Dublin) and for letting me give a very spontaneous and unexpected "elevator pitch" on my demo at Interspeech 2017. Their interest in my work, but most of all their proximity, has fueled my engines, especially during the last stage of writing this dissertation.

Last but not least, this dissertation was possible thanks to the financial support of the MULTISENSOR and KRISTINA projects, which have received funding from the *European Union's Horizon 2020 Research and Innovation Programme*. It has also been partly supported by the DTIC doctoral fellowship and the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme.

To conclude, I must admit that I am uncertain of what the future will bring and it gives me a sort of dizziness to think about it right now. But I am positive seasickness eventually comes to a halt. So, I am ready to spread out the sail and navigate unexplored oceans. Before that, it is time for the reader to sail across this dissertation. Be welcome on board!

Abstract

This dissertation presents an empirical study on the information structure–prosody interface based on: (i) a formal description of hierarchical thematicity within a systematic language model for natural language generation within the Meaning-Text Theory; and (ii) two approaches to prosodic representation, namely, ToBI labels and acoustic parameters. A corpus of read speech by twelve native speakers of American English is used to test the viability to approach prosody generation in synthesized speech from a communicative perspective. To this end, statistical, classification and implementation experiments are carried out. The main technical contribution consists in two tools: a data-driven module for thematicity-based prosody enrichment in a speech synthesizer and an automatic prosody tagger developed under an extension of Praat for feature annotation. Results prove that thematicity spans have distinct prosodic characteristics, as previously suggested in theoretical studies, and that a tripartite hierarchical thematicity is a more appropriate representation of information structure than traditional binary flat thematicity approaches for its integration in speech technologies.

Keywords: *information structure, communicative structure, thematicity, theme, rheme, prosody, ToBI, acoustic parameters, speech synthesis, TTS, text-to-speech, CTS, concept-to-speech, automatic prosody labeling*

Resum

Aquesta tesi presenta un estudi empíric de la interfície d'estructura informativa–prosòdia basat en: (i) una descripció formal de la tematicitat jerárquica que s'enmarca en un model del llenguatje sistemàtique per a la generació automatic del llenguatje natural dins del marc de la Teoria Sentit-Text; i (ii) dues representaciós prosòdicas utilizant etiquetas ToBI i paràmetres acústics. Es fa servir un corpus de parla llegida per doze parlants natius d'anglés americà per a comprobar la valideça de la generació de prosodia en síntesi de veu a partir de característicasas comunicativas. Amb aquest objectiu, es presentan experiments estadístiques, de clasificació i de implementació. La principal contribució técnica consisteix en dues eines: un mòdul basat en dades per el enriquiment prosòdic de parla sintetizada a partir de la tematicitat i un etiquetador automàtique de prosòdia implementat en una extensió de Praat per la anotació de característicasas lling´uísticas. Els resultats demostran que els segmentes de tematicitat estan caracteritzats per trets prosòdics específics, com s'havia sugerit en estudis teòrics previs i que la tematicitat tripartita jeràrquica és una representació més adecuada de la estructura informativa que las propostas tradicionals de tematicitat binaria plana anteriors per a la integració en las tecnologies de la parla.

Paraulas clau: *estructura informativa, estructura comunicativa, tematicitat, tema, rema, prosòdia, ToBI, paràmetres acústics, síntesi de veu, text-a-veu, concepte-a-veu, etiquetatje automàtic de prosòdia*

Resumen

Esta tesis presenta un estudio empírico de la interfaz estructura informativa–prosodia basado en: (i) una descripción formal de la tematicidad jerárquica que se enmarca en un modelo del lenguaje sistemático para la generación automática del lenguaje natural dentro del marco de la Teoría Sentido-Texto; y (ii) dos representaciones prosódicas utilizando etiquetas ToBI y parámetros acústicos. Se emplea un corpus de habla leída por doce hablantes nativos de inglés americano para comprobar la validez de la generación de prosodia en síntesis de voz a partir de características comunicativas. Con este objectivo, se presentan experimentos estadísticos, de clasificación y de implementación. La principal contribución técnica consiste en dos herramientas: un módulo basado en datos para el enriquecimiento prosódico de voz sintética a partir de la tematicidad y un etiquetador automático de prosodia implementado en una extensión de Praat para la anotación de características lingüísticas. Los resultados demuestran que los segmentos de tematicidad están caracterizados por rasgos prosódicos distintivos, como se había sugerido en estudios teóricos previos y que la tematicidad tripartita jerárquica es una representación más adecuada de la estructura informativa que las anteriores propuestas tradicionales de tematicidad binaria plana para su integración en las tecnologías del habla.

Palabras clave: *estructura informativa, estructura comunicativa, tematicidad, tema, rema, prosodia, ToBI, parámetros acústicos, síntesis de voz, texto a habla, concepto a habla, etiquetado automático de prosodia*

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

> ”Words, like little buckets, are assumed to pick up their loads of meaning in one person’s mind, carry them across the intervening space, and dump them into the mind of another.”
>
> — Charles Osgood

Osgood’s bucket metaphor describes the complex process of human communication as a simple machinery (Osgood, 1960). A bucket (or rather, a word) loaded with meaning from the mind of a speaker is dumped into the mind of an addressee. In this metaphor, the assumption that communication merely involves the activity of passing meaning from source to target without any other intention on the side of the speaker is put at stake. So it is quite inevitable to wonder whether a speaker just dumps content in the mind of the addressee, or rather, the speaker purposefully structures content in manageable chunks of information.

This dissertation departs from the assumption that information in human communication is conveniently organized, and that prosody is one of the means to express this structure, reflecting the communicative intention of the speaker. What is more, speech synthesis is argued to regard prosody generation as a simple machinery, similar in some respects to the bucket metaphor, which limits the expressiveness of synthesized speech and its integration in more complex conversational scenarios. To overcome these shortcomings, the study of communicatively–oriented prosody generation in speech synthesis is instrumental.

In this dissertation, I explore how the “information structure” (aka “communicative structure”) codifies the communicative packaging by means of prosody in human and synthesized speech. The integration of prosodic mechanisms that speakers use to guide addressees through the content of an utterance is hypoth-

esized to render a more appropriate synthesized speech. Apart from theoretical approaches to the so-called "information structure–prosody interface" and minor attempts to test basic notions of 'given' and 'new' information in text-to-speech applications, there are no previous works that address the integration of a versatile communicative model for prosody generation in speech synthesis.

Such a challenging enterprise encompasses the study of distinct subareas within linguistics, namely, information structure and prosody, and computer science, in particular, speech synthesis. These subareas conform extensive research areas on their own, and to provide an overview of each of them is out of the scope of the present dissertation. Instead, the focus is put on the empirical analysis of the information structure–prosody correspondence and on the integration of the findings in speech synthesis applications.

But, what is prosody exactly? And why is it not a trivial task to generate expressive prosody in speech synthesis? According to Hiroya Fujisaki:

> "Prosody is the systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production. Its realization involves both segmental and suprasegmental features of speech, and serves to convey not only linguistic information, but also paralinguistic and non-linguistic information." Fusijaki (2012)

Fujisaki's definition underlines some key ideas on the role of prosody, namely:

- prosody is a system whose main function is to organize (or structure) speech;
- prosody conveys different types of information, i.e., linguistic and non-linguistic information, also connected to the emotions conveyed by speech.

Even though linguistic studies underline the role of prosody in communication, the generation of expressive prosody in speech synthesis, in particular, in text-to-speech (TTS) applications, is far from being considered a solved problem. In the early development stages of TTS, the role of prosody was restricted to undertake some linguistic tasks (usually the simplest). Still nowadays, prosody generation focuses on the distinction between stressed and unstressed syllables, the prediction of prominent words within a sentence, the insertion of pauses and variations of fundamental frequency to distinguish, e.g., a statement from a question. These tasks are usually solved using a set of pre-determined rules based on basic textual cues such as punctuation marks and word order. Even though such a basic role of prosody contributes to make synthesized speech intelligible, there

is a much wider range of communicative functions that involve prosody (e.g., intentions, emotions, expressiveness), which incipient speech technologies ignored and state-of-the-art applications still do not fully address.

The lack of concern of the communicative functions of prosody has implications that may affect the understanding of speech. For instance, the misplacement of prominence within a sentence may result in a different interpretation of the utterance depending on which word is made salient. In the example taken from (Hirschberg, 2008), *Why don't you move to California?*, the utterance can be understood differently: as a simple question or as a suggestion, depending on which word carries prosodic prominence, i.e.: *why* for simple question and *California* for a suggestion. In the case of phrasing, punctuation usually (but not always) helps to distinguish between different syntactic constructions. In the example taken from (Price et al., 1991), *Mary knows many languages(,) you know*, a pause after *languages* will change the syntactic (and prosodic) structure of the utterance. These examples illustrate the importance of linguistically adequate prosody in communication.

It cannot be denied that important advances have been made in synthesized voice quality, especially with the development of machine learning techniques like neural networks (Watts et al., 2016) for speech signal processing; but once the intelligibility problem has been solved, at least to a certain extent, it is high time to tackle the more complex subtleties of human spoken communication. The following section elaborates on the reasons why such a communicative perspective is instrumental in a steady evolution of speech synthesis and its application in human-computer interaction technologies.

1.1. Motivation

Speech technologies have evolved in a relatively short time-span from undertaking mere reading tasks, as, e.g., the well-known MITalk (Holmes, 1987), to handling conversations with human interlocutors, for instance, as applications in health care (Bierner, 1998; Wanner et al., 2017). Virtual and robotic assistants are gaining impact in society despite the fact that they still perform a basic set of actions. This increase in impact is motivated by the ease of use of such applications that are controlled through voice commands instead of mouse or keyboard input and that produce speech within a dialog interaction instead of text. However, the shift from what is known as 'text-to-speech' (TTS) to 'concept-to-speech' (CTS) (Schweitzer et al., 2006) has not fully been accomplished yet, and one of the pend-

ing issues concerns the generation of natural prosody.

Even though state-of-the-art prosody modeling considers linguistic functions to a certain extent, prosody derivation in TTS applications is based on rules using low-level linguistic features such as word position, content-function word type, punctuation, etc.; see, e.g., (Olaszy and Nemeth, 1997; Tsai et al., 2014); on decision trees that draw upon a set of high-level linguistic features such as part of speech, dependencies, etc.; see, e.g. (Lindstrom et al., 1996; Xydas et al., 2005); or on a superimposition of prosody tags using a specific convention; see, e.g., (Delmonte and Tripodi, 2015). The role of prosody in structuring speech is reduced to a shallow function that permits basic intelligibility of the synthesized speech.

For instance, in TTS, punctuation is regarded as the only superficial textual marker for pauses. However, the connection between punctuation and prosody in textual and spoken natural language is not as simple as it may seem at a first glance, as pointed out by Shriberg (2005). Spoken language is structured according to the communicative intention of the speaker, and punctuation of a text only reflects a (rather small) subset of this structure. Moreover, punctuation is related not only to prosodic organization, but also to syntactic organization and sentence modality, among other functions, which do not always involve a pause. Despite this observation, TTS applications directly translate textual punctuation marks, such as commas and full stops to pauses in speech without considering other functions of punctuation (unrelated to pauses) and other dimensions of prosody (apart from phrasing). This is one of the reasons why, even though voice quality varies from system to system, prosody is often regarded as monotonous or even inappropriate, in some cases, mainly due to the fact that it lacks the communicative power and versatility present in human speech.

A monotonous prosody of TTS applications especially affects virtual social agents, also known as embodied conversational agents (ECAs): one of the biggest trials of artificial intelligence and a great step forward in human-computer interaction technologies. The challenge of computers being able to understand and react appropriately to human speech increases in difficulty depending on the tasks to be performed. Thus, CTS applications are expected not only to efficiently convey meaning, but also to engage the listener in conversation. For this reason, the synthesized speech needs to be communicatively expressive and natural, and thus, take the development of ECAs to a next level. This task requires an important update of the research agenda to include communicative approaches for prosody generation in CTS applications, since CTS currently deploys the same type of prosody modeling as TTS applications, ignoring what has been argued in

the literature for quite some time, namely:

- prosody expresses the communicative intention of the speaker (Grice, 1989);
- the communicative intention of the speaker is to a large extent encoded in terms of the information structure (Steedman, 2013);
- information structure is rendered both through syntax and prosody (Mel'čuk, 2001);
- in CTS, the information structure of a sentence can be derived in a content organization procedure, as done in natural language text generation (NLTG) (Wanner et al., 2003; Bouayad-Agha et al., 2012).

Should this argumentation hold, monotonous and unnatural prosody (especially in multiple sentence discourse) inherited from TTS technologies can be avoided (or at least reduced) in CTS applications by drawing upon the information structure derived automatically using techniques employed in natural language text generation (NLTG).

Despite the great interest in the study of the information structure–prosody interface in the linguistic, prosodic and computational research (see (Brown, 1983; Sgall, 2000; Büring, 2016), among others), communicatively-motivated approaches to prosody generation still remain largely unexplored, especially in implementation settings. Research and development of TTS applications (in particular for commercial use) have mainly focused on the development of high-quality voices and signal processing techniques, setting aside linguistically-oriented approaches. Minor attempts were made some time ago to integrate the information structure–prosody interface in speech technologies (Steedman, 2000; Kruijff-Korbayová et al., 2003; Haji-Abdolhosseini and Müller, 2003). It was usually one aspect of information structure that has been studied: thematicity. Thematicity defines how content is packaged in terms of "what is being talked about", i.e., the 'theme' and "what is being said", i.e., the 'rheme'. Most of the approaches drew upon such a binary flat thematic division and established a one-to-one correspondence between theme–rheme and rising–falling intonation patterns respectively. But two key issues were underestimated for an adequate integration of the information structure–prosody interface in TTS applications: the assignation of thematicity and the generation of a varied range of prosodic cues in the speech signal.

To determine theme and rheme in a statement, it is common to picture the statement as an answer (A) to a hypothetical question (Q), as the following example taken from (Steedman, 2000) shows:

(1)

(Q): I know what Marcel SOLD to HARRY.
But what did he GIVE to FRED?
(A): (Marcel GAVE)$_{Theme}$ *(a BOOK)*$_{Rheme}$ *(to FRED.)*$_{Theme}$

This example serves to clarify the concepts of theme and rheme, but it is difficult to apply to other contexts (e.g., to monologue speech). (Q) clearly states what information the speaker knows, i.e., *I know what Marcel sold to Mary* and what new information he requests *what did he GIVE to FRED?* (A) contains the new information (Q) requests, which is conveniently labeled as rheme, i.e., *a book*. Such a question–answer scenario, although sufficient for a laboratory experiment, is distant from spontaneous dialog settings (where, usually, known information is not made explicit) and monologues (where discourse evolves around statements). Moreover, approaches of this kind fail to provide a formal representation of thematicity that is instrumental in a CTS application.

Regarding prosody, theoretical studies on the information structure-prosody interface often refer to rising and falling intonation using the most popular convention in the area of speech prosody: the **To**nes and **B**reaks **I**ndex (ToBI) (Silverman et al., 1992). ToBI uses a symbolic alphabet for representing prosody contours. Tone variations are described as low ('L') and high ('H'), diacritics signal prominent syllables (i.e., '*') and phrasing (i.e., '%' or '–'). Thus, rising patterns associated to themes are represented as L+H* LH%. Going back to example (1), the L+H* LH% ToBI pattern associated to the theme *Marcel gave* indicates that there is a rising tone on the stressed last syllable of Mar**cel** and a rising boundary tone on the word *gave*. However, despite the popularity of the ToBI convention in theoretical studies on the information structure–prosody interface, the current implementation of ToBI labels in TTS applications is rather limited and supported only by some TTS applications.

A key issue concerning the implementation of ToBI in TTS is the one-to-one mapping of ToBI labels (e.g., H*) to acoustic parameters (e.g., an increase of 50% in fundamental frequency). This fixed mapping implies that every time a specific label is inserted, the same effect is produced on the synthesized speech. However, the ToBI convention establishes a contextual framework of reference to label prosodic events that may involve different types and degrees of prominence and phrasing.

Summing up, previous approaches to integrate the information structure–prosody interface in speech technologies have several drawbacks that motivate the need to

pursue a more appropriate methodology:

- no formal description of information structure is proposed;
- long sentences with complex syntactic structures are under-described in a binary representation of thematicity;
- a fully deterministic mapping between intonation labels and acoustic parameters is presupposed.

The present dissertation is prompted by the demand to address expressive prosody generation in CTS applications from a communicatively-oriented perspective and, thus, to give prosody in synthetic speech the credit it has in human speech concerning its function to structure the content of a sentence. Specifically, this dissertation addresses the aforementioned shortcomings and serves as a proof of concept for further development of the integration of communicative approaches in speech technologies.

1.2. Objectives

Despite the fact that some efforts have been made to generate more expressive prosody, to the best of my knowledge, there is no TTS/CTS application that can fully cope with the complexities of communicatively-oriented prosody generation. This dissertation addresses the following objectives:

- to advance in empirical studies for the computational analysis of the information structure–prosody interface, starting with the thematicity–prosody correspondence;
- to explore automatic approaches to prosody representation and provide tools that facilitate the analysis of large amounts of corpora;
- to implement prosodic modifications based on thematicity within a CTS application;
- to assess the appropriateness of such a thematicity-based synthetic prosody with quantitative and qualitative metrics.

The main objective of this dissertation is, therefore, to test the viability of a communicatively-oriented prosody generation in synthesized speech based on corpus analysis. Besides, automatic approaches for prosody annotation are tested to promote further advances in the study of large corpora that can be annotated semi-automatically. A final goal is to establish a methodology that serves to advance in the study of the information structure–prosody interface and, in parallel,

is deployable in an implementation setting on a CTS application.

Even though the present dissertation is restricted to a specific communicative aspect: thematicity, the overall methodology is designed to be scalable to other dimensions of information structure for their inclusion in later stages of development. Consequently, this dissertation aims at bringing forward communicative aspects to the research agenda within the field of speech technologies and, thus, foster further development in this direction. The following general hypotheses are tested:

- prosody and thematicity are related;
- different thematicity spans involve distinct prosodic cues;
- different speakers of the same language have a similar tendency to signal thematicity;
- there is a homogeneous distribution of acoustic parameters in specific thematicity spans across speakers;
- synthesized speech mimicking the thematicity–prosody correspondence found in human speech is perceived as more expressive than the default synthesis.

1.3. Outline of the Thesis

The thesis is structured in seven chapters.

- Chapter 2 includes the fundamental concepts regarding speech prosody and the information (or communicative) structure as defined by Mel'čuk (2001). An overview of Mel'čuk's theoretical approach to the communicative organization of natural language is presented to frame the object of study of this dissertation, i.e., hierarchical thematicity. Concerning speech prosody, a very brief overview of general concepts is introduced with a focus on the description of the ToBI convention. Then, a description of prosody enrichment in speech synthesizers is addressed; in particular, the conventions used in the experiments are presented.

- Related work is described in Chapter 3, including theoretical approaches to the information structure–prosody interface, implementation in TTS applications, and a brief summary of tools for automatic annotation and processing of speech prosody.

- The methodology regarding corpus compilation and experimental setup is introduced in Chapter 4. This chapter includes the description of the working corpus, and the procedure followed to annotate prosody. Finally, the

nature of experiments for the analysis of the information structure–prosody correspondence in human and synthesized speech is introduced.

- Chapter 5 outlines experiments concerning automatic prosody annotation. The chapter presents experiments using an existing tool for ToBI labeling, AuToBI. Besides, an experimental implementation of a rule-based approach to the segmentation of speech in prosodic units is introduced. This implementation involves the development of a modular automatic prosody tagger running on a web platform that includes an extension of Praat for feature annotation.

- The empirical analysis on the information structure–prosody interface is presented in Chapter 6. Experiments include hypothesis testing on the working corpus of read speech in American English using statistical and machine learning techniques. This chapter provides insights on several issues. Firstly, the working corpus is analyzed to observe what prosodic features signaling thematicity are shared among speakers. Then, binary flat thematicity is compared to tripartite hierarchical thematicity in their capability to predict prosodic cues. Afterwards, prosody prediction using high-level linguistics features that include hierarchical thematicity is explored. After that, the correspondence between hierarchical thematicity and prosody is described in terms of ToBI labels and normalized acoustic parameters. Finally, the implementation of a thematicity-based prosody enrichment in a CTS application is outlined and evaluated using perception tests and objective metrics.

- Finally, conclusions and future work are presented in Chapter 7.

Chapter 2

FUNDAMENTALS

> "A written sentence can be uttered in various ways to express different intentions, attitudes, and speaking styles which are under the conscious control of the speaker."
>
> — Hiroya Fujisaki

As already pointed out in the introduction, this dissertation encompasses the study of two research areas –namely, information structure and speech prosody– and its application in speech synthesis, which conform extensive scientific fields on their own. This chapter provides the fundamental notions from these areas that are essential to understand the core contribution of this dissertation. The integrative nature of the present study in the context of speech synthesis motivates some of the decisions that have been made throughout the development of the experiments, especially with respect to prosody representation.

The chapter is split into two sections. Section 2.1 briefly exemplifies terminological issues on information structure theories in general and, then, focuses on the representation proposed by Mel'čuk (2001) within the Meaning-Text Theory (MTT) (Mel'čuk, 1981). Mel'čuk's formal representation of the communicative organization in natural language is the corner stone of information structure, or rather, "communicative structure" (in Mel'čukian terms). The concept of thematicity in traditional and Mel'čukian perspectives is described. Afterwards, Section 2.2 sketches the main concepts around speech prosody relevant to this dissertation. The ToBI convention is outlined as the most widely used system for representing speech prosody and the one used in most studies on the information structure–prosody interface. Finally, prosody enrichment in TTS applications is introduced; in particular, the markup languages used in this dissertation to enrich prosody based on thematicity.

2.1. Information Structure

The way information is formally packaged in a sentence, known as "information structure", has been a fruitful field of research in linguistic studies to better understand how communication is produced and perceived. Information structure is a wide term and its study usually involves various linguistic dimensions in connection with how content is packaged, hence its interfaces: semantics, syntax and prosody, amongst the commonest ones. Studies use different terms to describe similar concepts like 'rheme' and 'focus'; cf., e.g., (Mel'čuk, 2001) and (Hajičova et al., 1998), among others. As already mentioned, the study of information structure in connection with prosody is often limited to thematicity in terms of what an utterance is about (the "theme") and what is being said about that (the "rheme").

The study of information structure is of relevance for natural language processing (NLP) since NLP is concerned with the creation of language models for computational tasks involving language understanding and generation, in particular for text generation. Still, only few NLTG applications actually include information structure in the process of generating text from ontological concepts. So far, only the representation proposed within the Meaning-Text Theory (MTT) by Mel'čuk (2001) has been (partially) deployed; in particular, his notions of communicative structure in (Wanner et al., 2003) and thematicity in (Ballesteros et al., 2015).

In this section, a brief introduction to theories on information structure is provided in Section 2.1.1. Emphasis is put on how information structure is studied in linguistics and some of the shortcomings of these approaches for NLP applications. Then, the ground concepts of the communicative structure by Mel'čuk within the MTT are introduced in Section 2.1.2. In Section 2.1.3, the thematicity proposed by Mel'čuk (2001) (which encompasses three thematicity spans and embeddedness) is compared to the traditional (flat binary theme–rheme) thematicity. Finally, guidelines for the annotation of text with Mel'čuk's tripartite hierarchical thematicity are presented.

2.1.1. Views on Information Structure

Although the literature speaks of information structure[1] in the context of studies on the correlation of prosody with the communicative intentions of the speaker, strictly speaking, it is just one dimension of information structure that has been

[1]Further reference on information structure theories can be found in the publication by Féry and Ishihara (2016).

taken into account so far: thematicity (or, in terms of the Prague School, *topic–focus articulation*). The key distinction within thematicity is the so called theme–rheme dichotomy; that is, a communicative segmentation of the meaning of an utterance into "what the utterance is about" and "what is being uttered about it".

The first theoretical studies on theme–rheme go back to Mathesius (1929), who seems to have borrowed the terminology from Ammann (1928). Later, the Prague school adopted the term *topic–focus* instead (Daneš, 1970; Hajicová, 1986; Sgall, 2000) to refer to this concept, and other authors from different schools of linguistics also did (Von Stechow, 1981; Lambrecht, 1994; Rooth, 1992). A number of other studies refer to thematicity with the term 'givenness'; see, e.g., (Schwarzschild, 1999), and thus talk about 'given' and 'new' information (Chafe and Li, 1976; Clark and Haviland, 1977; Brown, 1983). In other studies; see, e.g., (Firbas, 1964; Halliday, 1967; Mel'čuk, 2001), the terms "theme" and "rheme" persist.

Despite the great efforts along the years for defining these communicative notions, studies on information structure have remained within the field of theoretical linguistics. These studies sometimes explore different linguistic phenomena in relation to information structure (e.g., discourse, dialog, anaphora, and co-reference). The communicative structure within the Meaning-Text Theory comes to cope with some of the limitations other theories on information structure have, as this representation is devised in the context of a theoretical language production-oriented linguistic model, which is described in the next section.

2.1.2. Communicative Structure within the MTT

The Meaning-Text Theory (MTT) proposes a framework for language analysis and generation suitable for NLP applications. In particular, the "Communicative Organization of Natural Language" by Mel'čuk (2001) distinguishes different levels of representation. These levels are sequentially mapped from an unordered semantic representation (SemR) through a dependency tree structure of the Syntactic Representation (SyntR) and linearized chain of lexemes onto the Morphological Representation (MorphR) to get to the ordered string of phonemes at the Phonetic Representation (PhonR). Starting from SyntR and until PhonR, there is a subdivision into deep and surface representations.

The SemR is a predicate-argument structure and includes three components: the Semantic–Communicative Structure (SemCommS), which consists of a representation of the communicative intention of the speaker; the Rhetorical Structure, which encodes the artistic intentions and stylistic decisions of the speaker (irony,

humorous, etc.); and the Referential Structure, which specifies real-world referent for semantic configurations. The SemCommS superimposes on the SemR the communicative properties of the meaning of the sentence to be synthesized –rather than the communicative properties of the sentence itself[2]. Consequently, the functions of SemCommS are:

- organizing initial meaning into a message;
- ensuring coherence of the text of which the sentence under synthesis is supposed to be a part;
- reducing periphrastic potential of initial SemS, specifying more precisely the meaning.

In other words, the same abstract Semantic Structure can be shared by a given set of sentences, and by means of the SemCommS, these sentences are distinguished at subsequent levels (namely, SyntR, MorphR and PhonR). Figure 2.1 sketches the common SemS of sentences from (1a) to (1d) taken from (Mel'čuk, 2001).

(1a) *John met the doctor at the airport.*
(1b) *The doctor was met at the airport by John.*
(1c) *The airport was where John met the doctor.*
(1d) *It was John who met the doctor at the airport.*

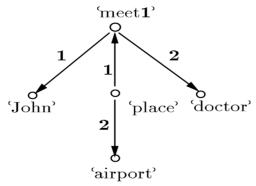


Figure 2.1: Shared SemS of examples (1a) to (1d) taken from (Mel'čuk, 2001).

The Deep Syntactic Structure (DSyntS) is the central component of the Deep-Syntactic Representation, which includes three further components: DSynt–Communicative Structure, DSynt–Anaphoric Structure and DSynt–Prosodic Structure (representing semantically conditioned prosodies). The DSyntS's of sentences (1a) (Figure 2.2) and (1d) (Figure 2.3) show how the Communicative Structure (CommS) determines different resulting dependency trees. The communicative

[2]In general linguistics, the term 'communicative' is usually linked to the idea of 'communicative competence' and refers to concepts related to the study of pragmatics; see the definition of 'linguistic competence' and 'performance' by Chomsky (1965).

subject (Theme) may coincide or not with the semantic subject (Actor) and syntactic subject (Synt-Subject), as represented in Table 2.1. This underlines the idea that CommS is a distinct dimension. Still, CommS may share elements with the SemR and/or the SyntR of a sentence.

Table 2.1: Communicative, semantic and syntactic subjects in examples 1a and 1d taken from (Mel'čuk, 2001).

(1a)	*John*	*met*	*the doctor*	*at the airport*
SemS	Actor			
SyntS	Synt-Subject			
CommS	Theme			
(1d)	*The doctor*	*was met*	*at the airport*	*by John*
SemS				Actor
SyntS	Synt-Subject			
CommS	Theme			

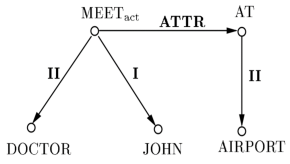


Figure 2.2: DSyntS from example (1a) taken from (Mel'čuk, 2001).

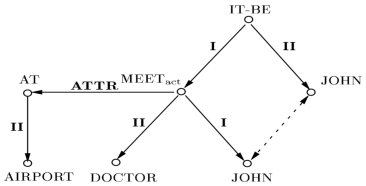


Figure 2.3: DSyntS from example (1d) taken from (Mel'čuk, 2001).

In a nutshell, CommS is part of the SemR and DSyntR of individual sentences. The communicative organization of text is not covered by CommS, it rather accounts for the structure of the so-called propositional content. Going back to example (1) taken from (Mel'čuk, 2001), the set of sentences may seem fully synonymous, but only (1a) is an appropriate reply to D1, whereas (1d) better suits D2:

D1 - *Nobody saw the doctor last night?*
- *John met him at the airport.*

D2 - *Ask John.*
- *Why John?*
- *It was John who met the doctor at the airport.*

CommS is composed of eight distinct dimensions: “thematicity”, “givenness”, “focalization”, “perspective”, “emphasis”, “presupposedness”, “unitariness” and “locutionality”. As CommS characterizes the meaning of the sentence and the sentence itself, it is, consequently, modeled at the semantic level, to be propagated then to the deep-syntactic and surface-syntactic levels of the linguistic description. Givenness, which is often treated as synonymous to thematicity, in Mel’čuk’s communicative structure theory is a distinct dimension from thematicity. According to Mel’čuk (2001), the thematization of the initial SemS has to do with psychologically motivated choices of the speaker, who decides that he wants to communicate some specific information (i.e., the rheme) concerning some specific item (i.e., the theme), and thereby makes the addressee follow him. In Mel’čuk’s words:

> “The Sem-Thematicity is thus a SPEAKER-ORIENTED Comm-category.” Mel’čuk (2001)

In sharp contrast, givenness has to do with the context-bound organization of the sentence, and in particular, with the addressee’s state of consciousness at the moment when the sentence is uttered. The concept of consciousness in speech had been previously addressed by Sgall (2000) and Chafe (1994), among others.

According to Mel’čuk, either theme or rheme may contain givenness. Givenness is independent, then, from thematicity structure and dependent on context and, more precisely, on the speaker’s knowledge of what information is shared by the addressee. In examples (2a-f), there are different sentences with the same SemCommS, as Table 2.2 adapted from (Mel’čuk, 2001) shows.

(2a) *Mary sprayed paint on a wall.*
(2b) *Mary sprayed the paint on a wall.*
(2c) *Mary sprayed paint on the wall.*
(2d) *Mary sprayed the paint on the wall.*
(2e) *The paint was sprayed on a wall.*
(2f) *Paint was sprayed on a wall.*

Sentences (2a) and (2f) do not contain any given items, but they indeed have a theme (i.e., *Mary* and *paint* respectively), as all the other sentences. Sentences (2b) to (2d) contain one or two given items in the rheme span (highlighted in bold in Table 2.2) and sentence (1e) has a given item in the theme span.

Consequently, the same given item, *the paint* (as in sentences 2b and 2e), can be part of either the rheme or the theme. There is a distinction between about what the speaker is saying something (e.g., the paint itself if that is the theme

Table 2.2: Thematicity and givenness in examples (2a) to (2f) taken from (Mel'čuk, 2001).

	Theme	Rheme
(2a)	*Mary*	*sprayed paint on a wall.*
(2b)	*Mary*	*sprayed* **the paint** *on a wall.*
(2c)	*Mary*	*sprayed paint on* **the wall.**
(2d)	*Mary*	*sprayed* **the paint** *on* **the wall.**
(2e)	**The paint**	*was sprayed on a wall.*
(2f)	*Paint*	*was sprayed on a wall.*

as in 2e) and, the fact that this is given information for the addressee. In other words, the shared context between speaker and addressee is regarded as a different dimension from thematicity, which establishes the speaker's communicative intention. This distinction between givenness and thematicity, together with the idea of focus, is often blurred in previous studies on information structure. As already mentioned above, in those studies, the terms 'given' and 'theme' are usually considered synonyms and not different dimensions, as Mel'čuk does. Such a distinction is instrumental for NLTG. In the following sections, I go into more detail of Mel'čuk's definition of thematicity, which is the dimension considered in this dissertation, and how it differs from traditional representations of thematicity.

2.1.3. Thematicity

There are two main views on how thematicity is defined. The first, traditional, view partitions a sentence into two subsequent flat spans, namely 'theme' and 'rheme', such that it is a binary division of a sentence related to both discourse and syntax layers (Erteschik-Shir, 2007). As Kruijff-Korbayová et al. (2003) explain referring to determination rules for information structure in a CTS implementation. To determine theme (T) and rheme (R) in a statement, it is common to picture the statement as an answer to a hypothetical question, as example (3) taken from (Steedman, 2000) shows. Thus, the theme is the part of the sentence that corresponds to what is being asked in the hypothetical question; and the rest is the informative part, which constitutes the rheme.

(3)
Q: *I know what Marcel SOLD to HARRY. But what did he GIVE to FRED?*
A: (*Marcel GAVE*)T (*a BOOK*)R (*to FRED.*)T

The second view is that advocated by I. Mel'čuk in the context of the MTT (Mel'čuk, 2001). Compared to the traditional theme–rheme dichotomy, thematicity in the MTT introduces two key features that enhance the scope of the theme–rheme span division, namely: (i) the notion of *specifier*, which sets up the context of the sentence, and (ii) the fact that thematicity is defined over propositions[3], rather than over sentences. This second feature implies that thematicity is *per se* hierarchical: if a proposition is embedded, its thematicity will be embedded as well.

Consider example (4), taken from our corpus, of the theme (T1) / rheme (R1) / specifier (SP1) distribution over propositions (P1, P2, etc.) in the sense of Mel'čuk:

(4)
{[*Ever since*]SP1, [*the remaining members*]T1 [*have been desperate for* {[*the United States*]T1(P2)[*to rejoin this dreadful group*]R1(P2)}P2]R1}P1

In example (4), the hierarchical thematicity structure is represented at different levels:

1. at level 1, P1 contains a theme, rheme and specifier;
2. at level 2, P2 is embedded into R1(P1) and has its own theme and rheme.

That is, spans at level 2 (as well as at subsequent levels) are hierarchically structured as an embedded thematicity representation; see Figure 2.4.

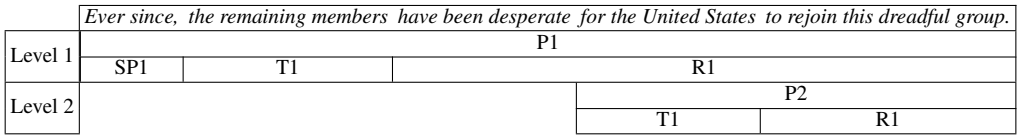


Figure 2.4: Hierarchical thematicity division of example (4).

As illustration, example (5) shows a theme–rheme segmentation captured in 'A(nswer)', as it is common in traditional studies of thematicity. Thus, a question (Q) is constructed to identify the theme (T), being the echo of the question, whereas the rheme (R) is the information provided to answer the question.

[3] A proposition is defined as the minimal syntactic unit with a conjugated verb and its dependent elements.

(5)
Q: *What happened ever since?*
A: [*Ever since*]T, [*the remaining members have been desperate for the United States to rejoin this dreadful group*]R.

Even though such a division is perfectly acceptable, it is totally dependent on the question that is being asked. Instead of asking "What happened ever since?", the question could be "What happened ever since to the remaining members?". In those cases, the theme span will extend to *Ever since* and *Ever since, the remaining members* respectively. Such methodology to establish binary thematicity seems right for a dialog interaction, but the lack of formal criteria to establish the theme–rheme division impedes the scalability to other genres such as monologues.

State-of-the-art theoretical approaches to the information structure–prosody interface (including Steedman (2000)'s among others) draw upon the first definition of thematicity in their study and so do implementations to TTS applications of this interface as it will be described in Chapter 3. However, to the best of our knowledge, no studies assessed so far the suitability of the Meaning-Text Theory (MTT) notion of thematicity with respect to its relationship with prosody and its application to speech synthesis scenarios.

The experiments presented in this dissertation in Chapter 6 demonstrate that Mel'čuk's definition of the hierarchical thematicity allows for a more accurate and fine-grained description of the relationship between how language is packaged and the prosodic events that contribute to this packaging. In what follows, a detailed description of annotation guidelines for hierarchical thematicity is introduced.

Hierarchical Thematicity

As mentioned above, the fact that thematicity is defined over propositions rather than sentences implies that thematicity is *per se* hierarchical, allows embeddedness and, thus, involves different levels of thematicity. For instance, a theme can be embedded in another theme or rheme span. A reference to the span that these embedded labels belong to is always written between brackets. For example, a level 2 T1 embedded in a level 1 T1 would be annotated as T1(T1) and a level 2 T1 embedded in a proposition would be labeled as T1(P2). Figure 2.4 shows the levels of embeddedness in example (4), where T1(P2), for instance, is a level 2 theme that is embedded in a level 2 proposition (P2). P2 is furthermore embedded in the main R1 span. As more than one thematicity span may exist within the same proposition, abbreviations include a number (e.g., 'SP1') that indicates the number of occurrences at each level (e.g., 'SP2' would be the second

specifier in a specific thematicity level).

	No one has worked out the players' average age, but most appear to be in their late 30s.				
level 1	P2		P3		
	T1	R1	SP1	T1	R1

Figure 2.5: Thematicity division in coordinated propositions.

In sentences containing coordinated propositions, there is a parallel thematicity structure (one partition by proposition) at level 1. In those cases, P1 is assumed (and not labeled) as the proposition containing the coordination and the two partitions are labeled as P2 and P3 respectively with a thematicity division each at level 1, as Figure 2.5 shows. The fact that thematicity in Mel'čuk's terms is called tripartite hierarchical thematicity in this dissertation does not mean that a sentence must compulsorily contain several levels of embeddedness nor three thematicity spans. As will be explained in the section on the annotation guidelines of thematicity in texts, a sentence may also be, e.g., rhematic, meaning that it only contains one rheme span.

Annotation Guidelines for Hierarchical Thematicity

The guidelines for annotation of hierarchical thematicity in text were defined and tested in (Bohnet et al., 2013). Propositions are the first units to be annotated, and theme and rheme constitute the communicative core (CC) of a sentence. Propositions may be a full clause (which contains a finite verb) or a reduced clause (where the corresponding finite verb is elided). Coordination and juxtaposition are annotated at level 1 (L1), in which case the first clause is annotated as P2, and subsequent clauses with correlative numbers, as P1 is considered to be the proposition containing all those coordinated or juxtaposed propositions.

All propositions (except for titles, which are all thematic) must at least contain a rheme. A rheme is what is being said about something; it is often recognized through exclusion or if it complies with the following characteristics:

- rhemes can be negated and/or questioned;
- existential clauses (those that begin with "there is/are") are all rhematic;
- non-fronted temporal, locative and manner circumstantials form part of the Rheme: [*I*]T1 [*met John some months ago in the park, in a very unexpected way*]R1.

As a theme is the text span about what the rheme says something, it should answer the question: *what about "the rheme"?* (where the rheme is substituted

by the words in the proposition). It can neither be negated nor questioned. In a relative clause, which is treated as an independent proposition, the relative pronoun is the theme only if it is subject (otherwise, it is a focalized part of the rheme). The following types of constructions comply with the characteristics of themes:

- titles are all thematic;
- in complex clauses with subordination where propositions are at the same level, the first proposition is all theme by convention;
- in adjectival constructions like *it's nice to see you*, the second part of the construction (i.e., *to see you*) is the theme;
- indefinite pronouns such as *nobody*, *somebody*, *nothing*, etc. and negative noun phrases cannot be themes: e.g., in *None of the boys did it*, it is not *none of the boys*, which is the theme, but rather *it*.

Specifiers do not express a separate message, but, rather, the context of the message to which they belong. Specifiers are annotated following these characteristics:

- fronted temporal, locative and manner circumstantials, e.g., {[*Apparently*]SP1 [*he*]T1 [*did so*]R1};
- fronted adjectival propositions with a sentential scope, e.g., {[*Tired of the same*]SP1, [*he*]T1 [*gave up*]R1};
- fronted discourse markers, e.g., {[*But*]SP1 [*it*]T1 [*was neither deep*]R1};
- circumstantials of the type "according to" (independently of their position), e.g., {[*About 25 % of the insiders*]T , [*according to SEC figures*]SP1, [*file their reports late*]R1};
- phrases that introduce direct speech (independently of their position), e.g., {[*It*]T1 [*is done*]R1, [*he said*]SP1};
- noun phrases in vocative case (independently of their position), e.g., {[*Anna*]SP1, [*he*]T1 [*did it*]R1}.

Such a thematicity annotation schema provides guidelines to annotate any type of text with hierarchical thematicity. The main contribution of this approach is that:

- it is systematic and language independent;
- it serves as a formal representation that is tested in the implementation of a communicative parser;
- it is oriented towards synthesis of a text instead of the analytic perspective taken by traditional approaches, which assign theme and rheme based on questions in a dialog setting;

- it is scalable to CTS transition-based parsing according to results in (Bohnet et al., 2013).

The working corpus presented in Chapter 4 and included as Appendix A was annotated following these guidelines. Annotated sentences are also made available as Appendix B.

2.2. Speech Prosody

Studies on speech prosody often suffer from what Xu (2011) calls *lack of reference*. He explains:

«By reference I mean a pivot that serves as both a starting point of inquest and a point that one can comfortably fall back on.» Xu (2011)

In his review on methodologies in the field of speech prosody, Xu points out the fact that methodologies are usually motivated by the kind of analysis (by transcription, by introspection, by hypothesis testing and by modeling) carried out and that they are also dependent on the aim to derive "descriptive" or "predictive" knowledge about prosody. Notwithstanding that agreement on a universal convention for representing prosody is currently one of the most important hurdles within the speech prosody community, proposals, such as (Hualde and Prieto, 2016), for an International Prosodic Alphabet are pushing the debate to come to an end. In any case, it is important to make clear what the starting point or *referent*, as Xu names it, is with respect to prosody representation in this dissertation.

The representations of prosody chosen for the present dissertation are motivated by their potential to modify prosody contours in TTS applications: (i) ToBI labels (Silverman et al., 1992), which uses discrete symbols to represent prosodic contours; and (ii) acoustic parameters, which can be used to specify changes in prosody contours in TTS by means of, e.g., the Speech Synthesis Markup Language (SSML) recommendation[4] (Taylor and Isard, 1997). This section sketches the main concepts around speech prosody used in this dissertation. A brief overview on prosody is presented in Section 2.2.1; the ToBI convention is described in Section 2.2.2; finally, Section 2.2.3 describes the prosody enrichment of synthesized speech that has been tested.

[4]SSML is not a prosody representation, but rather a convention to apply post processing modifications on synthesized speech.

2.2.1. Overview on Prosody

As already mentioned in the introduction, Fujisaki's definition of prosody implies a system whose main function is to organize (or structure) speech and it conveys different types of linguistic and non-linguistic information. He also proposes a hierarchical organization for prosodic units into: 'prosodic sentence', 'prosodic clause', 'prosodic phrase', and 'prosodic word'. These prosodic units involve several acoustic parameters and, according to Fujisaki, they do not always match syntactic constituents, especially in spontaneous speech.

Other authors, like Nespor and Vogel (1986) and Selkirk (1984), also address hierarchical prosodic structures in terms of constituents and their relation with syntax. These studies propose a division of prosodic units into 'utterance' (U), 'intonational phrase' (I), 'phonological phrase' (P), 'clitic group' (C) and 'word' (W), which is equivalent to the syntactic representation of a sentence. Figures 2.6 and 2.7 (taken from (Nespor and Vogel, 1986)) depict the correspondence between prosodic and syntactic structure. These studies highlight the role of prosody for the disambiguation of syntax.

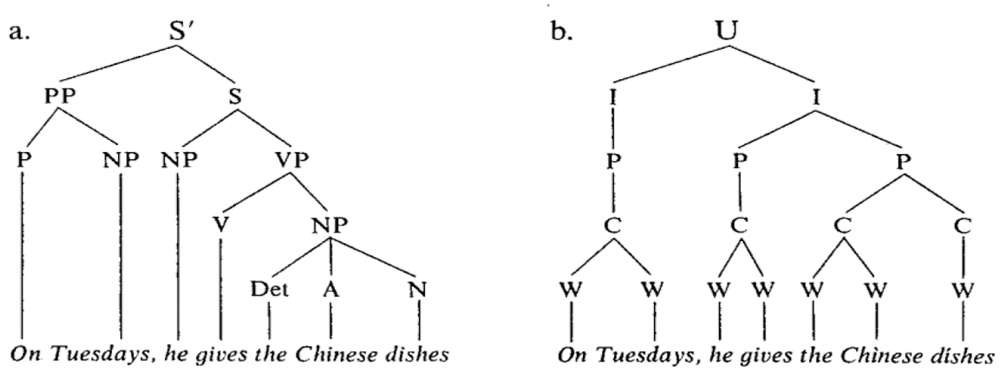


Figure 2.6: Phonological representation of prosody for DO taken from (Nespor and Vogel, 1986).

In these figures, the prosody-syntax correspondence leads to two possible configurations depending on whether the verb *give* is transitive (i.e., takes a direct object) or ditransitive (i.e., takes two objects: direct and indirect). In Figure 2.6, for instance, the direct object (DO)[5] of *gives* is *the Chinese dishes*, whereas in Figure 2.7, *the Chinese* is the indirect object (IO) and *dishes* the direct object. Consequently, different phonological phrases are formed namely, *he gives* (P1) and *the*

[5]The authors apply the term dependency on a constituency structure, i.e., the DO is a nominal phrase whose head is DO of the verb.

Chinese dishes (P2) in Figure 2.6 and *he gives the Chinese* (P1) and *dishes* (P2) in Figure 2.7. The difference in P1 and P2 disambiguates the syntax of these examples.

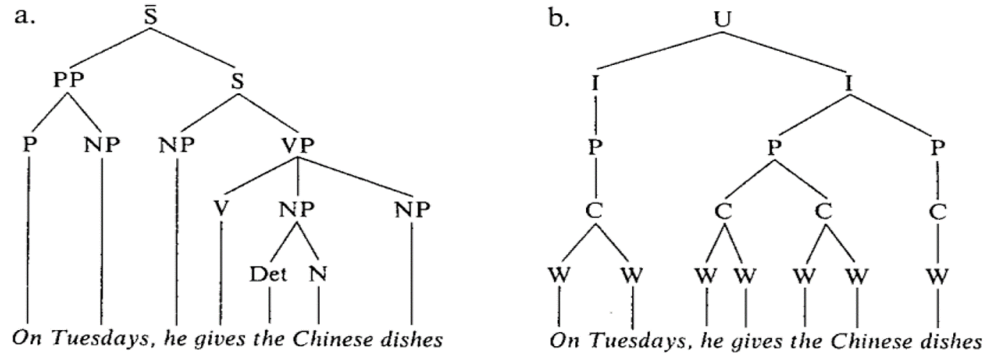


Figure 2.7: Phonological representation of prosody for IO + DO taken from (Nespor and Vogel, 1986).

For many authors, see, e.g., (Gussenhoven, 1984; Beckman and Pierrehumbert, 1986; Ladd, 1996), prosody primarily refers to the melodic nature of speech, i.e., intonation, strictly speaking, whose perception correlate is the variation of pitch, and the acoustic dimension is the variation of fundamental frequency (F0). These studies also acknowledge the function of intonation to segment speech, organized into constituents, that is, they talk about phrasing, and prominence within those phrases.

However, recent studies underline the role of intensity, see, e.g., (Tseng, 2004) (or loudness as its perception correlate) and rhythm, see, e.g., (Hirst, 2009), in prosody organization and modeling. Prosodic elements can be measured using a larger or more reduced set of acoustic parameters; see, e.g., (Tahon and Devillers, 2016). Some authors (Campbell and Mokhtari, 2003; Li et al., 2015) regard voice quality as part of prosody, especially connected to the study of spontaneous and affective speech.

In speech synthesis, it is quite common to address prosody modeling as the generation of F0 contours to convey linguistic information accounting for lexical stresses, prominence and phrasing; see, e.g., (Taylor, 1998; Anumanchipalli, 2013). The modeling of F0 contours generally follows a set of rules (or decision trees) that are mostly based on part-of-speech and punctuation information, such as, for example:

- content words (nouns and adjectives) are always stressed;
- accentuation of full verbs is preferred to modal verbs, which, in turn, is preferred to accentuation of adverbs;
- commas and full stops involve breaks.

A review on the progress to be made in this direction is found in, e.g., (Hirschberg, 2002; Mixdorff, 2002; Van Santen et al., 2013).

In this dissertation, prosody is regarded as a global term that involves several elements, whereas intonation is considered to be one aspect of prosody. Therefore, despite the fact that some authors treat the terms intonation and prosody as equivalent, prosody is used as the general concept that includes intonation as one of its elements. The prosodic phrase (PPh) is the chosen referent prosodic unit in the role of structuring language in connection to thematicity. Thus, prominence and phrasing are explored within the PPh using three prosodic elements, namely, intonation or (the variation of) F0, intensity and rhythm.

The idea of prosodic phrasing and prominence is instrumental for the naturalness of speech and plays an important role in the context of the "semantics–syntax–information structure". This importance led linguists and speech researchers to establish annotation standards for labeling, analyzing, computing and modeling prosodic cues. Some annotations encode prosodic information using a symbolic label alphabet, such as INTSINT (Hirst and Cristo, 1998), iViE (Grabe et al., 1998) and ToBI (Silverman et al., 1992; Beckman et al., 2004), among which ToBI is the most well-known and widely spread convention in the speech prosody community. Next section outlines the ToBI convention.

2.2.2. The ToBI Convention

The original ToBI convention (Silverman et al., 1992; Beckman et al., 2004) represents prominence and phrasing by means of discrete labels. According to the ToBI annotation guidelines[6], all words should be labeled in order to account for the "associated record of the fundamental frequency contour". The labeling is usually done in a TextGrid format using Praat, where 4 tiers are created, namely, an ortographic (for the transcript), a tone (for F0 tonal symbols), a break-index (for types of breaks) and a miscellaneous tier (for comments). Labels (in the tone tier) indicate F0 movements or variations along stressed syllables that are relative to the context of the sentences and cannot be described by a dictionary entry. Thus,

[6] https://www.ling.ohio-state.edu/∼tobi/ame_tobi/annotation_conventions.html; http://www.cs.columbia.edu/∼agus/tobi/labelling_guide_v3.pdf

prominence is associated to pitch accents (PA), which are identified by a star ('*') symbol; whereas phrasing is labeled as boundary tones (BT) signaled by a dash ('-') for intermediate phrases involving a break and a percentage symbol ('%') for intonational phrases usually involving a 'breath pause'. A letter or combination of letters is assigned to PA and BT, indicating that the F0 is high ('H') or low ('L'). If F0 movements involve a change in H or L within the same word, a bitonal label is assigned. Such an F0 shift may be ascending or descending and may occur before or after the stressed syllable. Thus, a post-nuclear (L*+H, H*+L) or pre-nuclear (L+H*, H+L*) rise or fall of F0 is annotated. In these bitonals, the plus symbol ('+') indicates a movement of F0 either within a syllable or across syllables. Breaks between words are labeled (in the break-index tier) with a number within a scale from '0' (when there is no break at all) to '4' (when there is full pause). Thus, an intermediate phrase involves a break of type 2 or 3 and an intonational phrase a type-4 break. In this dissertation, only tone labels are used, in line with previous theories on the correlation of rising intonation (L*+H LH%) to theme and falling intonation (H* LL%) to rheme spans. Figure 2.8 shows an example of standard ToBI annotation, where all words carry a label, as specified in the ToBI convention.

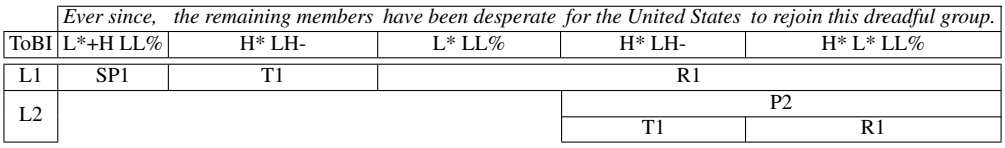


Figure 2.8: ToBI and thematicity annotation of example (4).

2.2.3. Prosody Enrichment in TTS Applications

Prosody enrichment in TTS applications consists in applying specific modifications on the default synthesized speech. Such prosody enrichment consists in specifying a certain modification of a particular word or group of words. Several XML-based markup languages are used to encode these modifications. The most well-known and pertinent to this dissertation are the *Affective Presentation Markup Language* (APML) (de Carolis et al., 2004) and the *Speech Synthesis Markup Language* (SSML) (Taylor and Isard, 1997).

APML and SSML instruct the TTS application to carry out appropriate actions in a standardized format. These conventions establish a way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. across different

synthesis-capable platforms. Markup languages establish a control sequence defined in terms of attributes that is mapped onto the acoustic signal and is, thus, parametric in nature. However, markup values are merely indications rather than absolutes. As stated in the SSML recommendation, for example, it is possible to explicitly indicate the duration of a text segment and also indicate an explicit duration for a subset of that text segment. But, if the two durations result in a text segment that the synthesis processor cannot reasonably render, the processor is permitted to modify the durations as needed to render the text segment.

In TTS applications, some attempts have been made to include ToBI labels (e.g., in the open-source speech synthesizers Festival (Black and Taylor, 1997) and MaryTTS (Schröder and Trouvain, 2003)) to manipulate the prosody contour using an XML-based markup language. The modification of prosody by means of ToBI labels, however, is restricted in some respects. For instance, the actual mapping of a ToBI label (e.g., H*) to the speech signal consists in assigning a fixed value of increase (e.g., 50%) in fundamental frequency (F0).

The SSML *prosody tag* allows control of six optional attributes: overall pitch, pitch contour, pitch range, speech rate, duration, and volume. These attributes can be modified independently or in combination. For our implementation, overall pitch and speech rate were chosen individually and in combination. Absolute (e.g., '+50 Hz' for increasing a specific amount of hertz (Hz) in F0) and relative values (e.g., '+20%' for increasing a percentage in F0) can be specified. An example of SSML prosody tag for modification of several prosodic attributes is presented in example (6).

(6)

<prosody rate="-10%" pitch="+20%">*Ever since,* </prosody>*the remaining members have been desperate for the United States to rejoin this dreadful group.*

The SSML *boundary tag* controls the introduction of pauses at a specific location. The duration of the break is specified in milliseconds (ms). Consider an example of an SSML boundary tag:

(7)

Ever since, the remaining members <boundary duration="100"/>*have been desperate for the United States to rejoin this dreadful group.*

Most of the theories on the information structure–prosody interface use ToBI to refer to rising and falling tunes associated to theme and rheme respectively,

so testing in implementation settings using these tags is useful to make a direct connection between theoretical and empirical approaches. On the other hand, a parametric approach (like the SSML-encoding) considers other acoustic parameters such as intensity and speech rate that are not directly represented in the ToBI annotation schema. Each approach has advantages and limitations that will be further explored in Chapter 6.

Chapter 3

RELATED WORK

> ”The basic idea behind all work in this area is that communication takes place against a background of shared knowledge so that the way a listener interprets an utterance will be partly dependent on the (situational) context in which the utterance occurs.”
>
> — Klaus von Heusinger

Most theories on information structure are oriented to the analysis of language from a theoretical linguistic perspective rather than to analysis and generation in a computational scenario. Something similar occurs when revisiting the literature on the information structure–prosody interface: most studies take the ‘analysis by introspection’ approach that results in a lack of formalism. A formal description (as in (Mel’čuk, 2001)) is a *sine qua non* requirement for the integration of the information structure–prosody interface in speech synthesis. However, no previous studies have explored the correspondence between hierarchical thematicity and prosody for its application to computational linguistics settings. In this chapter, I summarize previous work on other descriptions of information structure in connection to prosody from theoretical and implementation perspectives. Moreover, I introduce some tools for speech processing and automatic labeling of prosody that are relevant to the scope of the present dissertation.

Section 3.1 includes a brief summary of linguistic studies on the information structure–prosody interface. Section 3.2 presents the application of some basic notions of information structure to prosody generation in TTS/CTS applications. Finally, Section 3.3 describes the most common open source tools for processing, annotation and automatic tagging of prosody in the speech community.

3.1. The Information Structure–Prosody Interface

The interest in the information structure–prosody correspondence applied to speech synthesis lies in the derivation of prosody that is communicatively oriented and more natural. Knowing the linguistic mechanisms involved in human communication is pertinent to the achievement of multifaceted speech technologies that can carry out more complex tasks linked to conversational settings. The information structure–prosody interface stands out as a solid ground for starting to build up such a communicative model in the computational field. Let us review the state of the art with respect to their characteristics and shortcomings in view of this problem.

In the introductory quote to this chapter, von Heusinger (1999) refers to the information structure-prosody correspondence as the study of how shared knowledge affects the interpretation of a message. And indeed, the role of information structure in comprehension of read and spoken speech has been reported for a long time in linguistic and cognitive sciences (Clark and Haviland, 1977; Bock et al., 1983; Fowler and Housum, 1987; van Donselaar and Lentz, 1994). Recent studies in German (Meurers et al., 2011) and Catalan (Vanrell et al., 2013) also show that characteristic intonation patterns that make a distinction between theme and rheme spans contribute to a better understanding of the message. But what does this correspondence consists in?

The relationship between information structure and intonation had been discussed even before ToBI was agreed upon as a convention to represent intonation cues. Beckman and Pierrehumbert (1986) suggest that the characteristic bitonals for theme and rheme are L*+H and H+L* respectively. Steedman (2000) proposes a question–answer setting for the identification of theme and rheme and builds upon Beckman's assumption to hypothesize on complete intonation patterns for theme and rheme. In example (1), the theme span *Marcel gave* contains a rising PA (L+H*) on *Marcel* and a rising BT (LH) on *gave*.

(1)

Q: *I know what Marcel SOLD to HARRY. But what did he GIVE to FRED?*

A:	(*Marcel*	*GAVE*)	(*a*	*BOOK*)	(*to*	*FRED.*)
	L+H*	LH		H*		LL%

The problem here is that, if the text that is segmented is not inserted in a dialog, questions may lead to different thematicity segmentations. Examples (2a) to (2h), taken from (Haji-Abdolhosseini and Müller, 2003) illustrates this drawback. For example, the question "who gave the book to Mary?" would result in a segmentation as in (2b), where *Jane* is the rheme. But, the question "what did Jane

do?" would lead to the same segmentation as in (2b), but, in this case *Jane* will be theme and the rest rheme.

(2a) [*Jane gave the book to Mary.*]
(2b) [*Jane*] [*gave the book to Mary.*]
(2c) [*Jane gave the book*] [*to Mary.*]
(2d) [*Jane gave*] [*the book*] [*to Mary.*]
(2e) [*Jane*] [*gave*] [*the book to Mary.*]
(2f) [*Jane gave*] [*the book to Mary.*]
(2g) [*Jane*] [*gave the book*] [*to Mary.*]
(2h) [*Jane*] [*gave*] [*the book*] [*to Mary.*]

Haji-Abdolhosseini and Müller (2003) suggest that intonation and phrasing is what actually determines the information structure, i.e., a rising intonation on *Jane* (L*+H LH%) in (2b) will correspond to the theme assignation, while a falling intonation on *Jane* (H* LL%) will indicate that *Jane* is the rheme. On the other hand, (2c) will also be characterized by a rising final tone (LH%) on *book*, indicating that the whole segment (i.e., *Jane gave the book*) is the theme.

Some attempts have been made on exploring additional aspects of prosody, apart from F0 contours, in connection with information structure usually with respect to the dimension of focus. These studies are, as a rule, restricted to one prosodic element in isolation; see, e.g. (Calhoun, 2010) on rhythm (or, rather, 'metrical structure', as the author defines it); (Xu, 1999) on F0 alignment and (Féry, 2013) on prominence and phrasing. The concept of 'prosodic focus' is often studied in opposition to givenness. Büring (2003), for instance, studies the difference between 'contrastive', 'broad' and 'narrow focus' within given and new information spans. Also Kügler et al. (2013) analyze intonation as early focus and post-focal givenness in the context of speech synthesis, as well as the role of duration in marking prosodic focus. Finally, it is worth mentioning the work by Zubizarreta (2016) on nuclear stress and information structure. However, these studies use only small sets of examples to illustrate their hypotheses.

With respect to empirical approaches to the information structure–prosody interface, studies on a corpus of more than two speakers are uncommon. The intonation of givenness, for instance, is studied by Baumann (2012) in German using one speaker. Baumann (2012) provides evidence that a range of pitch accent types can be mapped onto a gradient scale of givenness degrees, with the pitch height on the accented syllable being the determining factor. Féry and Kügler (2008) study the process of tonal scaling on a corpus of German consisting of eighteen

speakers, 2,277 sentences of the same syntactic structure with a varying number of constituents, word order and given–focus structure. According to Féry and Kügler (2008): "a given constituent has been mentioned in the question or the context introducing the target sentence" whereas 'focus' can be 'wide' (or 'all-new') and 'narrow' defined as follows:

> "In an all-new sentence, no element has been mentioned in the preceding context or was especially prominent in the common ground of the protagonists. A narrow focus was induced by a context asking explicitly for one or more arguments, or for the verb." Féry and Kügler (2008)

A further issue is the availability of spoken corpora annotated with a range of linguistic layers that include information structure and prosody. Regarding prosody, there are a number of available resources especially in English[1] among which the most popular is the Boston University Radio Speech Corpus (BURSC)[2]. But with respect to information structure, even though there is some literature on annotation guidelines, e.g., (Baumann et al., 2004), annotated available resources are few: as far as one can tell from the literature, there is one in German (Stede and Mamprin, 2016) and the one used for the present study in English and already used in (Bohnet et al., 2013). These resources are restricted to the annotation of thematicity. Other dimensions of information structure are pending to be included in annotated corpora.

As aforementioned, in Mel'čuk (2001)'s theory, focalization, givenness and thematicity are different dimensions of communicative structure. In this dissertation, the only dimension under consideration is thematicity. And indeed, this is just a first step. But, in order to achieve natural prosody, other interfaces must be looked at and the application in the computational field is far behind the linguistic knowledge in this respect. For instance, prosody is explored in connection to semantics (Büring, 2016), pragmatics (Hirschberg, 2008), and syntax (Price et al., 1991), among other studies on prosody interfaces.

[1] see https://corplinguistics.wordpress.com/2012/03/08/prosodically-annotated-corpora/ for further references.

[2] https://catalog.ldc.upenn.edu/LDC96S36https://catalog.ldc.upenn.edu/LDC96S36

3.2. Application of the Information Structure–Prosody Interface in Speech Synthesis

Some minor attempts to include thematicity in TTS applications were made especially around the change of the century, like (Prevost, 1996). His work proposes a constrained based assignation of intonation (illustrated in Figure 3.1) derived from thematicity based on questions and semantic constrains to deal with words that carry contrastive focus. In a similar way, Haji-Abdolhosseini and Müller (2003) assign intonation using lists of words that match the 'given information' category, that is, the theme, in my terminology. Further examples of a system keeping track of information that has been previously mentioned in the discourse are found in (Kruijff-Korbayová et al., 2003; Kügler et al., 2012).

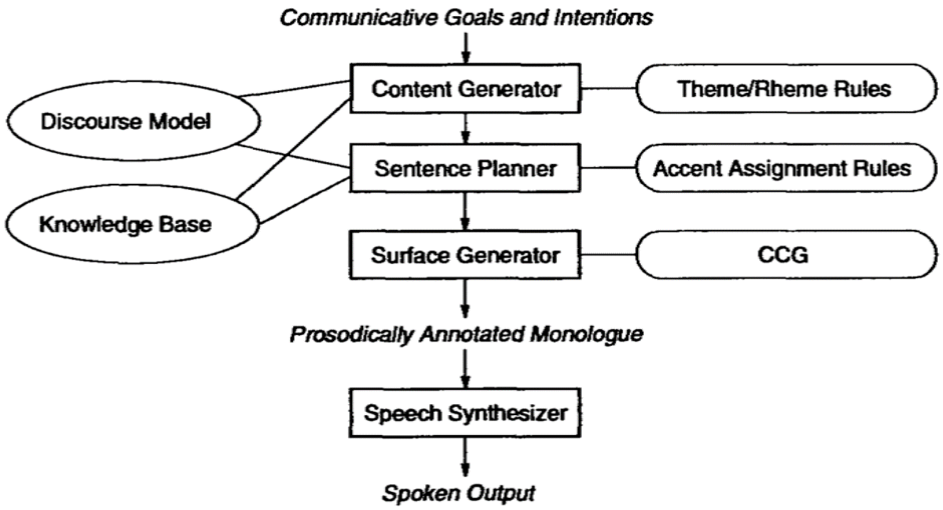


Figure 3.1: Architecture proposed by Prevost (1996).

Steedman (2000)'s study on the correlation of theme and rheme to rising and falling intonation patterns was extended for implementation using the APML convention to specify intonation (Steedman, 2004). Moreover, the Festival Speech Synthesizer's (Black and Taylor, 1997) prosody module includes a basic decision tree to assign rising intonation to themes (usually when the theme coincides with the subject of a sentence). The integration of Steedman's view of thematicity and the architecture proposed in (Prevost, 1996) is carried out by Kruijff-Korbayová et al. (2003) in a dialog system, which is tested using both FestivalTTS and MaryTTS.

MaryTTS (Schröder and Trouvain, 2003) is actively maintained and developed, what has probably led to researchers, especially in German (as this was the original development language), to continue implementing and testing the information structure–prosody correlation on this open-source TTS application. This has resulted in the creation of dedicated tags and implementation of intonation in MaryTTS for the notions of givenness and contrast. Specifications for MaryTTS implementation are given in (Romanelli et al., 2001). The authors first describe the general architecture of the system as shown in Figure 3.2, taken from (Romanelli et al., 2001). As can be seen in that diagram, prosody is specified at the same level as lexicon and letter-to-sound rules, but then duration and F0 generation occur in a final stage.

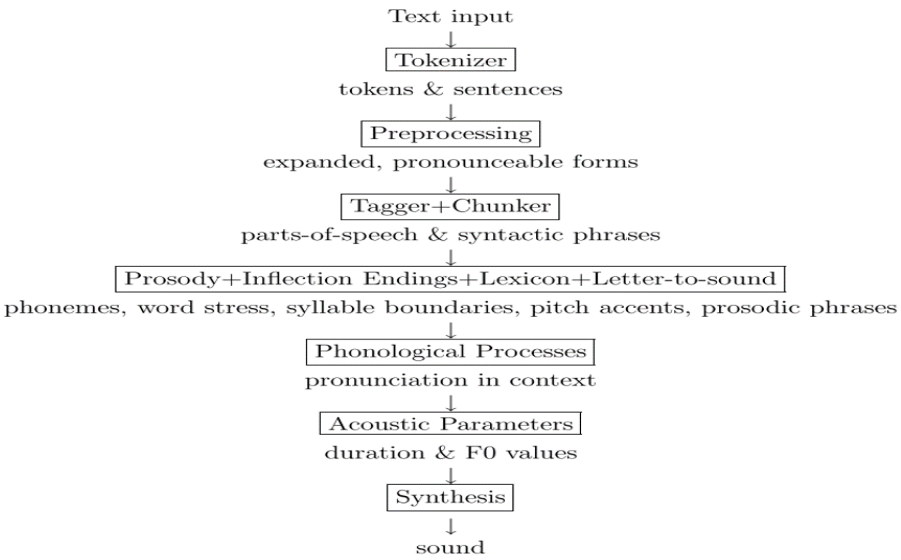


Figure 3.2: MaryTTS Architecture taken from (Romanelli et al., 2001).

The assignation of information structure status is based upon the recognition of words that have already been mentioned in the discourse (i.e., repetitions of lexical items) that are included in a 'GivenList'. Thereupon, Romanelli et al. (2001) propose that given information is unaccented whereas new information is accented to provoke a contrast. Besides, they consider that a word must bear a contrastive accent if that word is an antonym or a hyponym of a precedent word. In order to identify these words, negations and contrast markers are used as indicators. Such a definition of information status is an attempt to find a way to implement some basic notions of given and new information in a computational scenario based on intuition. In fact, this proposal of a given item being unaccented is an opposed

idea to theoretical studies saying that themes carry a rising intonation (and are, therefore, prosodically marked). Therefore, it cannot be said that this given–new notions are equivalent to theme–rheme according to the terminology used in this dissertation.

The aforementioned implementations of the information structure–prosody interface suffer from some deficiencies. The first problem is the strong contrast between the complexity in theoretical constructs and the simplistic application in implementation settings: for instance, to consider as given any word that is repeated in a text implies an obvious oversight of studies on information structure. These deficiencies, together with the lack of empirical evidence and formal representations of thematicity, have relegated linguistic approaches to prosody implementation to the background. The present dissertation explores a new approach in this respect with hierarchical thematicity as the basis to bring the information structure–prosody interface back to the research agenda in speech technologies. Given that a key aspect in the transition from linguistic studies to computational applications is the availability of large amounts of annotate corpora, the next section presents existing open source tools that permit processing and annotation of speech prosody for compilation of speech corpora.

3.3. Speech Processing and Annotation Tools

Annotated speech corpora are the starting point for training algorithms in computational applications. High-level linguistic annotation tasks involve a considerable amount of manual work by trained experts in the field. This has led to an increasing interest in automatic or semi-automatic tools for making the process more efficient and development of annotation software with visualization and scripting functionalities to automate routines or rule-based processes.

The Praat software (Boersma and Weenink, 2017) is the most well-known speech processing and annotation tool in the speech community. A brief overview of this software is outlined in Section 3.3.1 so as to underline some aspects that led to the development of the extension for feature annotation introduced in Chapter 6. Then, a brief summary on automatic labeling of prosody is presented in Section 3.3.2, focusing on existing tools.[3]

[3]The literature on theoretical aspects regarding automatic labeling considerably exceeds the number of available open source software.

3.3.1. Annotation and Scripting under Praat

Praat (Boersma, 2001; Boersma and Weenink, 2017) is an open-source platform for phonetic research used in the speech community for annotation, analysis and synthesis purposes. Praat is a powerful, user-friendly, programmable, freely available, and actively maintained software.

The strong point of Praat is its built-in speech processing functionalities that are accessible through a user interface. Styler (2013) provides a detailed description and guide on Praat for linguistic research. Other tutorials are available online[4]. Speech audio files can be visualized in their waveform[5] and spectrogram[6] representations (including formants, pulses, pitch and intensity contours, etc.) just by opening an audio file in the main object menu (see Figure 3.3).[7]

Praat includes a dedicated format called *TextGrid* for the annotation of sound files that contains a minimum of one tier. Each tier is mapped to the whole timestamp of the associated sound file, and includes interval or point segment annotations. Interval and point segments may take an optional label; this label is the only information that can be included into any annotation. Since the labels cannot be extracted as objects in the main Praat window, no action can be scripted based upon labels.

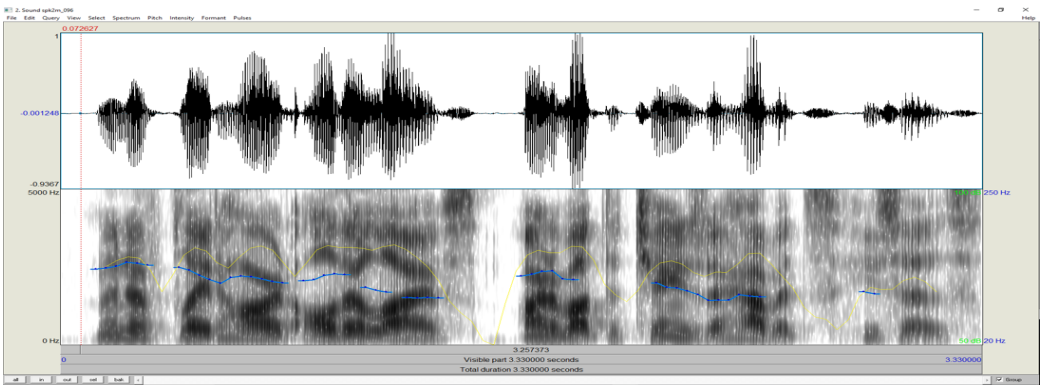


Figure 3.3: Waveform and spectrogram representation in Praat.

[4]https://web.stanford.edu/dept/linguistics/corpora/material/PRAAT_workshop_manual_v421.pdf and http://www.helsinki.fi/ lennes/vispp/lennes_palmse05.pdf

[5]A waveform is the curve showing the shape of the sound wave representation with respect to time.

[6]A spectrogram is the visual distribution of energy as a function of frequency for a particular speech sample.

[7]In Praat, the waveform is always located above and the spectrogram below.

Praat has a dedicated scripting language (based on the programming language C) and an environment to automatically run its multiple functionalities available from the object main menu (functionalities that are only accessible through the GUI cannot be scripted). Compilations of ready-made scripts are available from different webpages.[8]

While suitable for a coarse-grained glance at the acoustic profile of speech, Praat shows two major limitations when it comes to more detailed annotation that also involves linguistic information. Firstly, the segment annotations of Praat are opaque blocks of strings, and there is no function for a linguistic analysis of the labels. For instance, if an interval segment for a word, e.g., *places* includes morphological information within the same label (e.g., "places: noun = plural"), there is no function in Praat that would allow the division of the string "places: noun = plural" into tokens of any kind, for example, "places—noun—plural". Secondly, Praat is not modular, i.e., all automatic routines (e.g., detection of silent and voiced parts, annotation of intensity peaks and valleys, computing relative values, etc.) must be programmed together in a single script. No composition of stand-alone off-the-shelf scripts for dedicated subroutines is possible, which implies that for any new constellation of the subroutines a new script must be programmed.

In order to remedy these limitations, advanced users have found workarounds. Thus, the first limitation is solved by either extracting information to an external file, as ProsodyPro (Xu, 2013) does, or by annotating in parallel tiers with cloned time segments and different labels. To circumvent the second limitation, experienced users tend to program in external platforms and call Praat for performing specific speech processing routines. For example, Praaline (Christodoulides, 2014) extracts acoustic information from Praat for analysis in the R statistic package (R Core Team, 2013) and visualization in the Sonic visualizer (Cannam et al., 2010). However, these workarounds make the use of Praat cumbersome.

In order to address these limitations of Praat, the *Praat on the Web* tool will be introduced in Chapter 5. Praat on the Web upgrades Praat along the lines observed in state-of-the-art NLP annotation interfaces as encountered for SEMAFOR[9] (Tsatsaronis et al., 2012), Brat[10] (Stenetorp et al., 2012), or GATE[11] (Cunningham et al., 2011, 2013). Such an upgrade is instrumental for the study of prosody interfaces and a versatile semi-automatic approach to annotation and a

[8] https://sites.google.com/site/praatscripts/, http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html and https://lennes.github.io/spect/

[9] http://www.cs.cmu.edu/ ark/SEMAFOR/

[10] http://brat.nlplab.org/

[11] https://gate.ac.uk/

compact visualization of those features is essential for the integration of linguistic interfaces with prosody.

3.3.2. Automatic Prosody Labeling

An increasing interest in automated prosody labeling was experienced the decade following the introduction of the ToBI convention, mainly to avoid the time-consuming procedure of manual annotation. Thereupon, a rather extensive number of works focused their interest on the automatic detection, modeling and annotation of prosodic events in speech; see, among others:

- the Fujisaki model, (Hirose et al., 1984; Mixdorff, 2015; Salvo Rossi et al., 2002);
- the INTSINT representation, (Hirst, 2001; Hirst and Auran, 2005);
- the Common Prosody Platform (CPP)[12] (Prom-On et al., 2016): an open initiative for comparison of F0 models, namely, the *Command-Response* (CR) model, the *Autosegmental-Metrical* (AM) model, the *Task-Dynamic* (TD) model and the *Target Approximation* (TA) model;
- the annotation of ToBI for different languages including Japanese (Noguchi et al., 1999), Korean (Lee et al., 2002), Spanish and Catalan (Elvira-García et al., 2016);
- annotation of prosody based on tonal perception (Mertens, 2004);
- annotation of prosodic phrases, e.g., ANALOR[13] (Avanzi et al., 2008) implemented for the MATLAB environment (MATLAB, 2007).

Regarding automatic annotation tools for prosody based on machine learning techniques, AuToBI[14] (Rosenberg, 2010) was the first publicly available tool to automatically annotate prosody (F0 contours and breaks) for American English with ToBI labels (Silverman et al., 1992). AuToBI is trained on an English corpus of broadcasting radio news, making it domain- and language-specific. AuToBI outputs word-by-word annotation. In line with AuToBI, ANALOR (Avanzi et al., 2008) is trained on a small corpus of radio broadcast in French. Like AuToBI, it is domain- and language-specific, and it allows segmentation of an utterance into major prosodic units. The issue on the availability of annotated corpora arises

[12] http://commonprosodyplatform.org/
[13] http://www.lattice.cnrs.fr/Analor.html?lang=fr
[14] http://eniac.cs.qc.cuny.edu/andrew/autobi/

again, as it would not be difficult to retrain these tools if large amounts of annotated corpora were available.

In the field of prosody annotation, it is common that the object of research usually determines the usability of the methodology and tools, as Batliner and Möbius (2005) highlight. This involves that testing needs to be carried out in order to assess the usability of existing tools for the intended task. Chapter 5 includes a detailed account of experiments carried out using two versions of AuToBI for labeling the working corpus of read speech in American English used to explore the information structure–prosody correspondence in Chapter 6.

Chapter 4

METHODOLOGY

> ”It is important to get results from experiments but
> the most important is the process in getting results.”
> — Dr. Nik Ahmad Nizam

This chapter introduces the methodology followed for the analysis of the information structure–prosody interface in this dissertation, focusing on how data was collected, processed and analyzed. This methodology lines up with the objective to provide empirical evidence on the information structure–prosody correspondence from corpus-based experiments.

The previous chapter instantiated the fact that most theoretical approaches that study the information structure–prosody interface suffer from substantial empirical evidence of their postulates. According to Xu (2011)'s review of the methodologies in the field of speech prosody, theoretical studies on the syntax-pragmatics-prosody interfaces use an “analysis by introspection” as methodological approach. He argues that this approach is imprecise as the assignment of prosody “by intuition” is inevitably unreliable. The present dissertation proposes an approach that tests previous theories, expands these theories by a formal representation of information structure and tests the obtained results in speech synthesis experiments. Moreover, experiments on automatic labeling of prosody are also carried out with existing tools, and new tools are developed to annotate and extract prosodic information from speech. The proposed methodology facilitates a steady transition towards a more solid empirical ground in this field of study and contributes to the integration of communicative approaches in speech technologies.

As stated before, previous studies of the information structure–prosody interface proposed *ad hoc* postulates, where the analysis of empirical data was ex-

ceptional and limited to short sentences in question-answer settings. These approaches have impeded steady progress, especially in the application to speech technologies. This dissertation envisages the study of the information structure–prosody interface from a methodological perspective based on empirical testing in two experimental setups: corpus analysis of human speech and speech synthesis experiments. The proposed methodology aims to achieve the following goals:

- to allow scalability to other communicative dimensions, registers and languages developing tools for the automatic annotation of prosody;
- to analyze the thematicity–prosody correspondence in human speech using a corpus-driven approach;
- to explore the advantages and limitations of a data-driven thematicity-based prosody enrichment in a CTS application.

Such a methodology addresses two main research issues in this field: (i) the lack of empirical analysis of the information structure–prosody correspondence; and (ii) testing of the integration of the information structure–prosody interface in computational settings. The proposed methodology involves a standard scientific flow of knowledge derived from each stage into the whole system as represented in Figure 4.1.

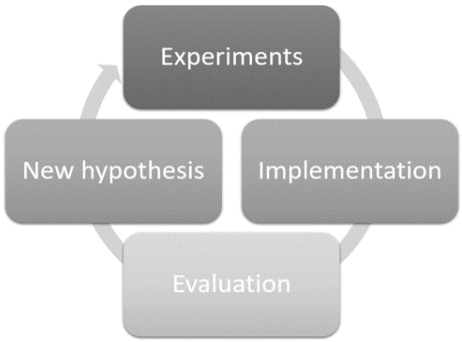


Figure 4.1: Methodology knowledge flow.

This chapter is structured as follows. Section 4.1 explains the criteria followed in the corpus compilation. Then, Section 4.2 outlines the methodology used in the manual annotation of prosody and the criteria established for the automatic extraction of prosodic parameters. Finally, Section 4.3 details the design of experiments introduced in Chapter 6.

4.1. Corpus Compilation

The rest of the section outlines the corpus characteristics from a textual and speech point of view together with the datasets that have been derived from the corpus.

4.1.1. Textual Characteristics

To set up the corpus, a selection of 109 isolated sentences from the Wall Street Journal (WSJ) was made. These sentences contain a variety of topics and a varied range of linguistic structures. The chosen text genre is news since the information structural component is expected to be richer in this type of written discourse. Sentences are extracted from different pieces of news, as the study analyzes the sentence as the referent linguistic unit. The WSJ Penn Treebank (Charniak and al., 2000) is annotated with other linguistic dimensions, such as part of speech (PoS) tags and syntactic relations (converted to dependency syntax relations), which are used in some of the classification experiments.

From the communicative perspective, a representative amount of hierarchical thematicity spans is chosen taking into account that there must be examples of:

- rhematic sentences, i.e., sentences that contain only a rheme;
- theme–rheme structures of different lengths (long sentences are preferred);
- a tripartite division into theme, rheme and specifier;
- different levels of embeddedness, at least level 1 (L1) and level 2 (L2);
- different syntactic sentential constructions: juxtaposed, coordinated and subordinated.

The corpus contains: simple sentences, coordination, subordination and the combination of both. This varied syntactic composition is related to the representativeness of communicative structure in terms of:

- the number of thematicity levels (up to three in the corpus);
- the position of spans within the sentence and with respect to each other;
- and the continuity or lack of continuity of spans (in particular, rheme spans can be discontinuous).

Figure 4.2 shows the distribution of the most frequent thematicity spans and multiple propositional sentences (P2) in the corpus on a relative scale of 0 to 1, where ‘0’ means that there is no example in the whole corpus, and ‘1’ there is an example in each sentence of the corpus. All sentences contain a rheme (R1),

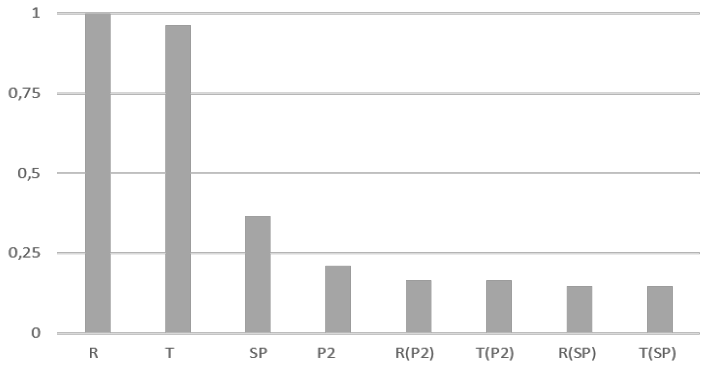


Figure 4.2: Distribution of the most representative thematicity partitions in our corpus.

96% have a theme (T1), and 37% include a specifier (SP1). 21% of the sentences contain more than 2 propositions (P2). Embedded spans within a level 1 specifier (labeled as T1(SP1), and R1(SP1)) are represented in 15% of the sentences. The rest of spans is represented in a proportion of 6% or below. These under-represented spans in the corpus include discontinuous rhemes (i.e., R1-1 and R1-2 labels), second specifiers (i.e., SP2), second rhemes (R2) and embedded spans in themes and rhemes (e.g., T1(T1) and R1(R1)).

The complexity in communicative structure is related to the number of propositions and embedded thematicity spans within each sentence. In this respect, 70% of the corpus contain L1 thematicity, 14% of the sentences contain more than one proposition (P2, P3, P4, etc., which are further subdivided into thematicity spans) and 16% of the sentences involve embedded spans (i.e., T1, R1 or SP1, which are subdivided at L2 into further thematicity spans). Most of L1 thematicity spans (11% of the total 16%) that contain embedded thematicity are specifiers, followed by rhemes (4%). Embeddedness in themes is rarely found in our corpus (only 1% of the total).

In terms of the number of words, the corpus has an average of fifteen words per sentence with a minimum of three words and a maximum of thirty. Figure 4.3 shows the distribution of sentences in quartiles with respect to their length in words. The highest concentration of sentences (54%, to be precise) is found in the third quartile, i.e., sentences containing between sixteen and twenty-three words.

Themes have been chosen for further insight in the analysis of the information structure–prosody correspondence. As previous work proved that the theme span

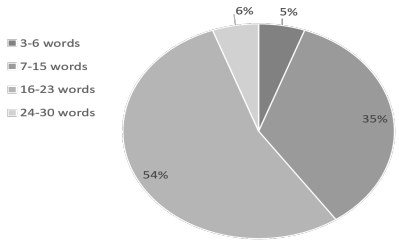


Figure 4.3: Distribution of sentences according to the number of words.

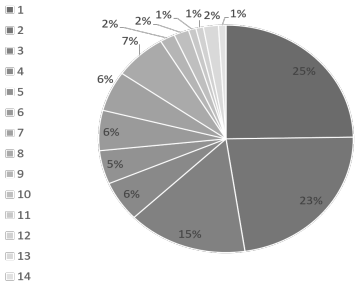


Figure 4.4: Distribution of themes according to the number of words.

is prosodically distinct (characterized by rising intonation patterns), this hypothesis is explored in connection to the number of words in the theme span. Thus, Figure 4.4 includes the distribution of themes according to their number of words. One-word themes represent 31% of the total number of themes (i.e., 152 themes including all levels of embeddedness), and 44% contain more than three words.

4.1.2. Speech Characteristics

Regarding the selection of participants, native speakers of American English were recruited. If previous studies on the information structure–prosody correspondence are proved to be right, parameters such as gender, age, cultural background and region of birth should not affect results of the study.

The corpus was recorded in a professional studio located in the facilities of the University Pompeu Fabra. While recording the corpus, participants were instructed to read naturally. Once participants read the corpus, they were asked to speak spontaneously for about 3 to 5 minutes about any topic of their choice. These spontaneous speech samples are not included in the analysis of the infor-

mation structure–prosody interface in this dissertation. However, some of these samples are used for the evaluation of the automatic prosody tagger presented in Chapter 5.

A total of fifteen people was recorded reading the corpus. Three participants were discarded from the analysis because they exhibited speech disfluencies affecting their prosody when reading the sentences. Therefore, samples from twelve speakers were finally included in the corpus.

There is a balanced six-to-six distribution of male and female speakers. Participants are assigned an anonymous identifier with the format: *speaker (abbreviated as 'spk') – number (a correlative natural number) – gender ('f' for female or 'm' for male)*, resulting in, e.g., 'spk1f'. The participants were born in different dialectal regions in the USA (see figure 4.5) and showed different foreign language influences: all of them had been exposed to, at least, one foreign language (European Spanish) and most of them were fluent in this language. The majority of the participants (seven out of fifteen) belong to the North-Midland dialectal area in the USA.

Figure 4.5: Dialectal origin of participants.

Participants were asked to make a short pause after each sentence, as the experiments are restricted to this linguistic unit. They were asked to read the sentences naturally, and they were instructed to take the initiative and repeat sentences if they felt the sentence had not sounded natural, a word had been mispronounced or words were grouped together awkwardly. Some sentences (numbered as in the Appendices A and B) contained low frequency words or long noun compounds (highlighted in bold) even for the journalistic discourse which made sentences hard to read, for example:

(51) *This is the U.N. group that managed to* **traduce** *its own charter of promoting education, science and culture.*

(61) *The Babelists of the United Nations are experts at* **obfuscation**.

(65) *They* **sow a row of male-fertile plants** *nearby, which then pollinate the male-sterile plants.*

(75) *What triggered the latest clash was a* **skirmish** *over the timing of a* **New Zealand government bond issue**.

(87) *But for the next few months,* **these boys of summers long past** *are going to be* **reveling** *in an Indian summer of the soul.*

Regarding punctuation, which is known to affect prosodic phrasing when reading (Kalbertodt et al., 2015), a representative number of punctuation marks was taken into account, as detailed in Table 4.1.

Table 4.1: Usage of punctuation included in the corpus.

Punct. mark	**Usage**	**Sent n.**	**Example**
Comma	enumeration of nouns	39	[...] *images of clouds, beaches, deserts, sunsets, etc.*
	separation in phrases	52	*Ever since, the remaining members* [...]
	separation in clauses	65	[...] *male-fertile plants nearby, which then pollinate the male-sterile plants*
Semi-colon	introducing an enumeration	37	*Mr. Stoltzman introduced his colleagues:* [...]
	juxtaposition	48	*But it was neither deep nor lasting: light entertainment that was* [...]
Quotes	direct speech	72	"There is a large market out there hungry for hybrid seeds," he said.
	proper names	39	*"Deep Peace" also featured a slide show* [...]
	emphasis	22	*The new "social choice" fund* [...]
Question mark	open question	46	*What's next?*
	yes–no question	41	*Was this why some of the audience departed before or during the second half?*
Other	backslash as comma	37	[...] *pianist/bassoonist/composer* [...]
	hyphen as semi-colon	50	[...] *organizations - UNESCO*
	long hyphen as semicolon	90	[...] *you want one more – one more at-bat*

4.1.3. Datasets for Classification Experiments

Thirteen datasets have been created from the corpus, as reported in Table 4.2. Datasets marked with an asterisk symbol '*' are reduced datasets for specific experiments, with a selection of some speakers. Reduced datasets starting with the abbreviation 'AL' are subsets of the full dataset ALD. The upper part of the table shows datasets that include ToBI annotations (from ALD to L2TD). Instances are words in all datasets that contain ToBI annotation.

Table 4.2: Datasets derived from the corpus of read speech.

Acronym	Dataset Name	Speakers	Attributes	Instances	Type of instance	Classes
ALD	Acoustic and Linguistic features Dataset	12	20	18,792	Words	9
ALRD *	Acoustic and Linguistic features Reduced Dataset	5	20	7,830	Words	9
AL1FD *	speaker 1 female dataset	1	20	1,566	Words	9
AL1MD *	speaker 1 male dataset	1	20	1,566	Words	9
AL2MD *	speaker 2 male dataset	1	20	1,566	Words	9
AL4MD *	speaker 4 male dataset	1	20	1,566	Words	9
AL5FD *	speaker 5 female dataset	1	20	1,566	Words	9
AL5FTD *	speaker 5 female standard ToBI annotation dataset	1	20	1,566	Words	28
L2TD	Linguistic features to ToBI Dataset	12	15	18,792	Words	9
TRD *	Theme–Rheme Dataset	2	10	420	Binary spans	2
HTD *	Hierarchical Thematicity Dataset	2	10	575	Tripartite spans per level	15
SSD	Sentence Span Dataset	12	11	1,308	Sentences	17
TSD	Thematicity Span Dataset	12	14	6,036	Hierarchical spans	31

The lower part of Table 4.2 refers to datasets that contain a parametric representation of prosody derived from the automatic extraction and computation of acoustic parameters from the corpus. Four datasets are created extracting acoustic parameters from different segments. The theme–rheme dataset (TRD) contains 420 instances and two distinct classes, namely, theme and rheme as proposed in the traditional segmentation of information structure found in the literature, e.g., (Steedman, 2000). The hierarchical thematicity dataset (HTD) includes 575 instances and a total of fifteen distinct classes of hierarchical thematicity, as proposed in (Mel'čuk, 2001). Acoustic data from all twelve speakers is included in the sentence and thematicity span dataset (abbreviated as SSD and TSD, respectively). The main difference between these two datasets is that in SSD the segments are sentences and the classes to be predicted account for the L1 thematicity of each sentence, whereas in TSD the segments are thematicity spans with their corresponding labels assigned to them.

4.1.4. Data Protection Issues

In the compilation of a corpus involving speech samples, personal data protection issues must be taken into account. Speech is considered as "sensitive data" by the European law (Article 29 of the EU Directive 959/46/EC)[1] because individuals can be identified by their voice, which implies a breach of their right to remain anonymous. Therefore, when a corpus of speech is compiled, participants must be

[1] http://ec.europa.eu/justice/data-protection/article-29/documentation/other-document/files/2011/2011._04_20_letter_artwp_mme_le_bail_directive_9546ec_annex1_en.pdf

informed that their data is used for scientific research within the field specifically addressed in the study. All information concerning how personal data will be used within the scope of the project must be conveniently structured in an "informed consent" that the participants sign. This document guarantees participants that researchers are using their sensitive data only for the intended research purpose and that they guarantee the participant's right to ask for deletion of their data from the corpus. A template informed consent is included as Appendix C.

For reproducibility purposes, processed datasets and anonymized material are made available in my repository[2]. Pitch and intensity objects extracted with Praat do not serve to reconstruct an individual's voice, and thus they are not considered sensitive personal data, in contrast to raw speech audio recordings. An identification number is assigned to the speech samples as a pseudonymization procedure. The textual corpus is included in Appendix A and the annotation of the corpus with thematicity in Appendix B. When examples using sentences from the corpus are indexed with correlative numbers, a reference to the corresponding sentence number in the appendices is included as a footnote.

4.2. Prosody Representation

Prosody representation is a key aspect in this dissertation, in particular, automatic approaches to the annotation of speech prosody. As the main goal is to analyze the information structure–prosody correspondence in human read speech, as well as to provide a data-driven approach for the generation of synthesized speech prosody, experiments are designed to test manual and automatic approaches to prosody annotation.

As previously mentioned in Chapter 3, there are few available tools to annotate speech prosody, and most of them are language dependent or require a manually annotated corpus of a considerable size. Experiments on automatic prosody labeling using AuToBI are introduced in Chapter 5 as well as a tool developed to advance in the area of automatic prosody annotation using a rule-based approach.

The methodology for prosody representation underlying the automatic prosody tagger experiments with the hypothesis that prosody involves three acoustic elements, namely, intonation (or variation of F0), (variation of) intensity, and rhythm. In this dissertation, the prosodic phrase is the referent unit, and thus the automatic prosody tagger segments both raw speech and speech with word alignment us-

[2]http://github.com/monikaUPF/

ing rules for the parametric representation of prominence and phrasing within the prosodic phrase.

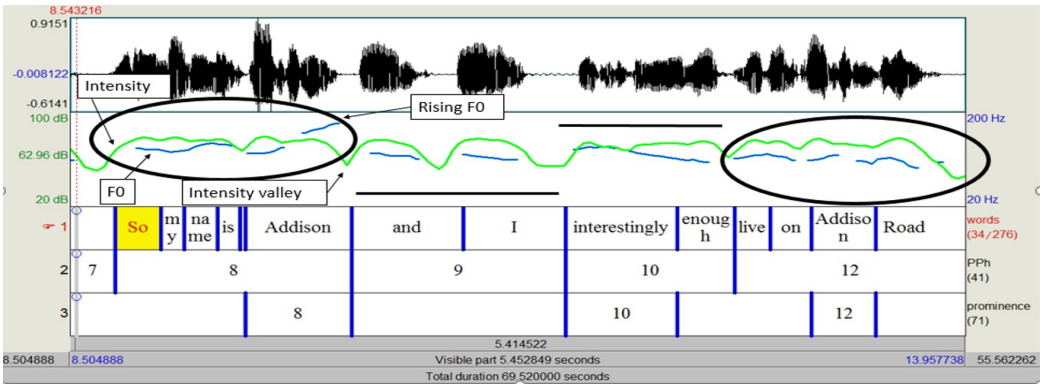


Figure 4.6: Example of prosodic units.

Figure 4.6 exemplifies graphically the parametric nature of prosodic phrases observed in the F0 (blue line) and intensity (green line) contours provided by Praat. These lines form a homogeneous picture (marked by an ellipse in Figure 4.6) that aligns with the PPh division. Nevertheless, there are some areas where the division is not that clear, as can be observed in the central part of the utterance. This structure observed in Figure 4.6 (which is also audibly perceived by an expert annotator) is translated into a vector of normalized acoustic values. Thereupon, prosodic units that are to be tagged correspond to positive or negative deviations of acoustic parameters.

Experiments on ToBI labels described in terms of a combination of normalized acoustic parameters are also carried out. However, as the implementation of thematicity-based prosody enrichment is carried out using acoustic parameters, further exploration of automatic tagging of ToBI labels is not pursued in the implementation of the prosody tagger.

4.2.1. Manual Annotation Criteria

Two different types of speech samples have been manually annotated: spontaneous and read speech. The annotation tasks that are to be carried out have a diverse nature: spontaneous speech is used for the experiments of rule-based automatic annotation of prosody; in particular, for the evaluation of the prosody tagger outlined in Chapter 5; whereas the corpus of read speech is used in the analysis of

the information structure–prosody correspondence presented in Chapter 6.

Consequently, this subsection includes two parts: Section 4.2.1 presents the guidelines for the annotation of prominence and boundaries of spontaneous speech at the PPh level; and Section 4.2.1 describes the adaptation of ToBI carried out for the annotation of read speech.

Annotation of Prominence and Phrasing

A set of guidelines is devised for the annotation of prominence and phrasing relying mainly on objective criteria that are based on the variation of acoustic parameters. Since further research on the spontaneous speech is foreseen as future work, and the spontaneous register has inherent difficulties especially for segmentation, specific notes on how to proceed in controversial points are included in the guidelines.

A PPh is defined as a prosodic entity that forms a homogeneous unit in terms of F0, intensity and duration cues and is signaled by one or a combination of acoustic parameters. The PPh is established as the immediately subsequent prosodic level smaller than (or, in some cases, equal to) the sentence. A PPh is marked according to the following criteria:

- In case there is one or (usually) a combination of the following conditions: pause, final rising intonation, lengthening of the last word, sharp fall in intensity, a PPh boundary is to be marked.

- In terms of content packaging, a PPh must contain at least one complete unit (usually a predicate with its arguments) notably large[3] in length with respect to the whole utterance and associated voiced segment respectively.

- In spontaneous speech, disfluencies such as disruptions, truncated phrases and hesitations may influence manual labeling of prosodic units. Therefore, all these events are to be included in the closest PPh to the right or to the left, depending on a pause preceding or following such disfluencies.

- If the contour following an unvoiced phoneme (with an *undefined* F0 value) is perceived as a continuation of the previous F0 contour (forming an homogeneous unit), no boundary is to be inserted. On the contrary, if the F0 contour is notably different after the F0 phonemic disruption compared to the preceding contour, a boundary is to be marked.

[3]It is left to the annotators' judgment what is considered 'notably large'.

Prominence within each PPh is marked in accordance with the following criteria:

- Prominent words are defined as a combination of one or (usually) several of the following parameters: F0 peak, high intensity, longer duration within its PPh.
- At least one word must be labeled as prominent within each PPh.
- Perceived relevant content must not be used as a criterion to label prosodic prominence (e.g., in noun compounds, an element tends to be perceived more prominent as it carries the semantic meaning of the unit).
- If a combination of acoustic parameters occurs within a word[4], for instance, increase in intensity and duration, this word should have more probabilities to be considered prominent than another word showing, e.g., an increase in F0 only.

Annotation using ToBI

The manual annotation of prosody using ToBI is carried out on words as referent units. Prosody contours concerning syllable segments are not considered and break indexes are not included in the annotation. Table 4.3 shows the inventory of ToBI labels used in the annotation. Words that are not prosodically marked ('False') and words that carry a prosodic label ('True') are annotated in each PPh. Then, words marked as 'False' are annotated as lexically stressed ('S') or unstressed ('U'), whereas words marked as 'True' are labeled as pitch accents (PA) or boundary tones (BT). Each PA and BT takes one of the possible ToBI labels shown in Table 4.3. With respect to bitonals, Hualde (2000) states that it remains to be demonstrated that L*+H and L+H* (which account for rising pre- and post-nuclear contours) have different pragmatic implications and, thus, can be considered phonological. In line with Hualde's view, the annotation of rising bitonals L*+H and L+H* converges to one label: L*+H[5]. A reduction of the ToBI catalog is carried out to simplify the number of possible assigned labels and accounts for a phonological description of prosody. Such a phonological representation of prosody is also in line with the implementation possibilities of ToBI labels within a TTS application, which are restricted to a limited realization of F0 contours for now.

[4]We refer to textual units in this case, as we are not aiming at segmenting prosodic words yet.

[5]Hualde proposed the label of (L+H)*, but such a label does not exist as prosody modification for TTS (that is usually done over words rather than syllables), hence the use of L*+H.

Table 4.3: Prosody annotation scheme for ToBI labels.

Prosodic Marker	Prosodic Type	Prosodic Label
True	PA	H*
		L*
		L*+H
	BT	HL%
		LL%
		LH%
False	S	
	U	

Concerning the function of prosody to signal the communicative elements within a hierarchical thematicity structure, the following aspects regarding prosody representation are considered to establish the correspondence between prosody and thematicity:

– Segment coincidence: full coincidence consists in a prosodic phrase containing a whole thematicity span. If a thematicity span partially coincides with a prosodic phrase, it means that the thematicity span contains more than one PPh, and the last PPh coincides with the end of span.

– Associated ToBI patterns: analysis of intonation tendencies across a variety of speakers.

– Acoustic parameters: analysis of distribution of mean values extracted at thematicity partitions including other prosodic elements apart from F0 prominence and phrasing, namely, intensity and rhythm.

4.2.2. Automatic Extraction of Prosodic Parameters

Automatic extraction and computation of acoustic parameters is carried out using the tools developed for experiments on automatic prosody annotation explained in Chapter 5, namely: the extension of Praat for feature annotation and the automatic prosody tagger. The modular nature of the prosody tagger allows the division of complex tasks in specialized subtasks that facilitate the processing, extraction and computation of acoustic parameters related to speech prosody used in the experiments on the correlation of information structure and prosody, especially in the implementation of thematicity-based prosody enrichment.

Table 4.4 shows the complete list of absolute and relative acoustic parameters (grouped by the three acoustic elements: F0, intensity, and rhythm), and abbrevi-

ations (within brackets) used in this dissertation.

Table 4.4: Prosodic elements and acoustic parameters used in this dissertation.

Element	Absolute Parameter	Relative Parameter
F0	mean F0 (F0)	z-score F0 (z_F0)
	standard deviation F0 (std.F0)	
	minimum F0 (min.F0)	
	maximum F0 (max.F0)	time point of max.F0 (maxF0.t)
Intensity	mean intensity (int)	z-score int (z_int)
	standard deviation intensity (std.int)	
	minimum intensity (min.int)	time point of min.Int (minInt.t)
	maximum intensity (max.int)	
Rhythm	duration (dur)	z-score dur (z_dur)
	speech rate in words/sec (sr.w)	z-score sr (z_sr)
	speech rate in syllables/sec (sr.s)	

Absolute values are extracted using different pre-determined functions available in Praat. Normalized values relative to the whole sample are computed for each segment of analysis, usually a thematicity span (it may be another segment, e.g., a word). Normalized values for mean absolute values of F0, intensity and speech rate are computed using the 'z-score' normalization. The z-score indicates how many standard deviations an element is from the mean. Z-scores are computed following the equation 4.1:

$$z_{score} = \frac{x - \mu}{\sigma} \tag{4.1}$$

where:

x = mean value of each acoustic parameter from a given thematicity span,
μ = mean value of the same acoustic parameter in the corresponding, and sentence
σ = standard deviation of the same acoustic parameter in the corresponding sentence.

Parameters referring to a time point are computed extracting the point of maximum F0 and minimum intensity respectively and calculating the relative time position in the span with a minmax score. Minmax normalization is computed following the equation 4.2:

$$minmax.t = \frac{x.t - min.t}{max.t - min.t} \tag{4.2}$$

where:

x.t = point in time where a peak or valley is located within an interval (e.g., word),
min.t = starting point in time of the corresponding interval, and
max.t = ending point in time of the corresponding interval.

In the minmax normalization, the minimum value is the starting time of the interval, which is mapped to 0, and the maximum value is the ending time of the interval, which is mapped to 1. So, the entire range of time points is mapped to the range 0 to 1. This gives us an idea of the relative time location of the peak within a time segment (in this case a word). In other words, the computed minmax score provides information on the location of the F0 peak ('maxF0.t') and intensity valley ('minint.t'). Thus, if an F0 peak is located within the first half of the time span, it will have a score between 0 and 0.5, and if an intensity valley is located within the second half of the span, the score will be between 0.5 and 1.

4.3. Nature of Experiments

Experiments in this dissertation are presented in two chapters: Chapter 5 unfolds around experiments on the automatic annotation of prosody, and Chapter 6 describes experiments on the information structure–prosody correspondence, including corpus-based and speech synthesis experiments. In what follows, the nature of these experiments is outlined.

4.3.1. Automatic Prosody Annotation Experiments

Automatic prosody annotation is essential for the development of large annotated resources. As previously mentioned, one of the biggest hurdles that hamper the integration of communicatively-oriented speech technologies is the lack of annotated corpora and open-source tools for the automatic annotation of linguistic resources.

Experiments on automatic prosody annotation are designed to test the performance of existing available tools, in particular AuToBI, on the working corpus. Then, the design of a rule-based tool for prosody annotation is devised to specifically meet our requirements, but also bearing in mind scalability for future work

and usability for other researchers. The experiment on developing this tool for automatic prosody tagging proves the hypothesis that a combination of rules based on acoustic parameters serves to automatically segment speech in prosodic units and detect prominence within these units.

4.3.2. Corpus-based Experiments

Corpus-based experiments are designed to both test the correspondence of information structure with prosody using ToBI labels and explore implementation possibilities in a TTS application. I acknowledge that prosody is influenced by other linguistic layers beyond information structure; in particular, by syntax and phonology. Still, due to the strong interaction between thematicity and prosody, it is legitimate to presuppose a bidirectional transition between them. A combination of statistical tests and machine learning techniques is used in the corpus-based experiments for hypothesis testing. The following assumptions are considered:

- a one to one relationship between prosody and thematicity is presupposed, acknowledging that they are both dependent on other linguistic dimensions;
- such relationship improves understanding of an utterance and is, thus, instrumental for TTS applications.

The software used for the statistical analysis is PSPP[6] (Pfaff, 2015) and Weka 3.8 Workbench[7] (Hall et al., 2009) for classification experiments. Classification experiments involve an unbalanced distribution in the number of attributes or classes (number of distinct labels to be predicted) in the working corpus. So as to compare results in this type of experiments, a majority voting ZeroR classifier is used as baseline (BL) and the improvement over the baseline is the metric to interpret the results. The absolute improvement (AbsImp) of the chosen classifier (CF) over this baseline is used as the assessment metric computed as:

$$AbsImp = \mu(CF) - \mu(BL) \tag{4.3}$$

where:

AbsImp = absolute improvement
μ(CF) = performance (i.e., accuracy, precision, recall, etc.) from the chosen classifier
μ(BL) = performance (i.e., accuracy, precision, recall, etc.) from baseline (ZeroR classifier)

[6] Available from: https://www.gnu.org/software/pspp/
[7] http://www.cs.waikato.ac.nz/ml/weka/downloading.html

The classification algorithm for ToBI prediction is a standard tree classifier (J48 in Weka) that is found in state-of-art prosody modules for speech synthesis (for example, in the module for prosody prediction in Festival TTS). In the case of testing with acoustic parameters, a bagging classifier with a REP tree is used.

As themes are proposed in previous theoretical studies to be distinct from rhemes in terms of intonation (rising versus falling patterns), a closer analysis in this direction is also deployed. Within this setup, the analysis of how the number of words affects the prosodic characterization of theme spans is carried out.

4.3.3. Speech Synthesis Experiments

Experiments on speech synthesis are envisaged as a test of viability of the implementation of a thematicity-based prosody enrichment in a CTS application. The proposed implementation of prosody enrichment makes use of the SSML convention[8]; in particular, the SSML *prosody tag* as specified for MaryTTS (Schröder and Trouvain, 2003).

Based on results obtained from the corpus analysis experiments, the characterization of hierarchical thematicity using z-scores is mapped onto a pair list of acoustic parameters associated to specific thematicity partitions in a selection of sentences representative of a variety of communicative structures. The resulting enriched synthesized speech is evaluated using a mean opinion score (MOS) test within a 1 to 5 Likert scale and a pairwise comparison for the subjective assessment of perception of expressiveness. Naturalness is assessed by means of objective metrics that quantitatively express the similarity of the resulting synthesized sample with a gold standard (human) speech sample.

[8]https://www.w3.org/TR/speech-synthesis/

Chapter 5

EXPERIMENTS ON AUTOMATIC PROSODY ANNOTATION

> ”In the case of prosody, very little of its functionality is orthographically represented, except for the punctuations whose meanings are at best ambiguous. Thus the starting point of inquest of prosody is inevitably vague and arbitrary, and it is just as difficult to know for certain what to check against after an observation is made.”
>
> — Yi Xu

Automatic annotation of speech prosody is instrumental for advancing in the development of more expressive speech synthesis. This chapter accounts for experiments related to the automatic annotation of ToBI labels using the open-source tool AuToBI and the development of our own framework for the annotation of prominence and phrasing based on a parametric approach: the automatic prosody tagger. This rule-based implementation serves to test the hypothesis of the three prosodic element correspondence of acoustic parameters for the segmentation of prosodic phrases and tagging of prominence within these segments. Moreover, the automatic prosody tagger is evaluated on spontaneous speech samples in English and Spanish to emphasize the scalability of this approach to the non-trivial task of segmenting spontaneous speech and the possibility of segmenting samples in different languages. The prosody tagger is designed as a modular script that uses an extension of Praat for feature annotation. The prosody tagger also serves to compute and create datasets with automatically extracted prosodic parameters that are used to analyze the thematicity–prosody correspondence in the next chapter.

This chapter is divided in three sections: Section 5.1 unfolds around experiments on automatic ToBI labeling using two versions of AuToBI; Section 5.2 describes the implementation of the functionality for feature annotation developed as an extension of Praat and the automatic prosody tagger; in Section 5.3, the chapter is finalized with a discussion.

5.1. Automatic ToBI Labeling

AuToBI is an open-source software[1] developed by Rosenberg (2010) to automatically label speech samples with ToBI labels that requires a TextGrid with the specific word alignment. As AuToBI is trained on a corpus of broadcast news, the initial expectation is that it should yield good results on our corpus, as it contains sentences from the Wall Street Journal, that is, a journalistic discourse as well. However, after testing two different versions of AuToBI, namely, the first version (v1.0) and the latest version available at the time of writing this dissertation (v1.5), results were suboptimal. In both tests, AuToBI's output had to be adapted to our requirements. The following sections explain how this adaptation has been done on each version and the results obtained.

5.1.1. Experiments on AuToBI v1.0

The first version of AuToBI annotates ToBI labels on all words. As our methodology considers only prominence and phrasing at the PPh level, a rule-based procedure for the adaptation of AuToBI's word-by-word output to PPh is devised. This adaptation also serves the goal to automatically establish a link between the information structure of an utterance and its prosody. The procedure groups AuToBI word labels into prosodic phrases and proposes a single intonation pattern for each PPh by means of a set of rules.

A selection of sentences from the corpus annotated with thematicity has been made. This further allows us to establish the correspondence between the automatic prosodic patterns and thematicity structure. This correspondence is used to validate the proposal of automatic prosodic pattern annotation assuming the classical work by (Steedman, 2000), which states that theme tends to be associated to the patterns L*+H and LH% (a rising intonation pattern), while rheme tends to be associated to the patterns H* and H*LL% (a falling intonation pattern).

The adaptation procedure consists of five different stages, as shown in Figure 5.1: (S1) thematicity annotation, (S2) corpus recording, (S3) AuToBI annotation,

[1] http://eniac.cs.qc.cuny.edu/andrew/autobi/

(S4) output adaptation, and (S5) validation of the results using manual reference annotations and the outcome of stage (S1). In the first stage (S1), the reference corpus is annotated with the information structure (focusing on the thematicity categories theme and rheme). In the second stage (S2), the reading of the corpus (or, as in our case, of a subset of the corpus) by one native speaker of American English is recorded. In the third stage (S3), the recorded speech is automatically labeled with the AuToBI tool. In the fourth (adaptation) stage (S4), the AuToBI word-by-word labels are transformed into PPh pattern labels. A final stage (S5) is used to assess the obtained patterns by comparing them with manual annotations and validate them with Steedman's theory on the correlation between prosody and theme/rheme structures. A description of stages 3 to 5 is provided below.

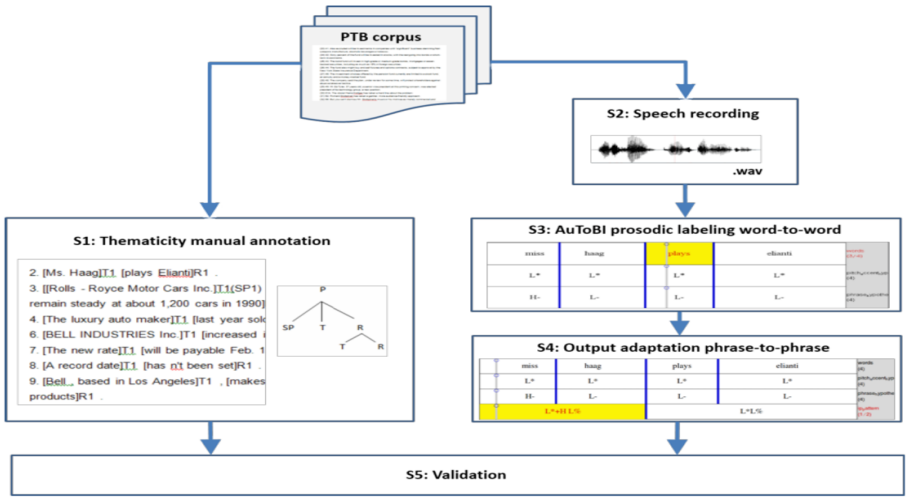


Figure 5.1: Processing pipeline on AuToBI_v1.0 output.

Automatic Prosodic Annotation Stage (S3)

This stage consisted in segmenting manually audio files into words as required by AuToBI. This has been done to automatically process AuToBI's labeling using Praat and thus generating a TextGrid file for each audio file. Results were saved as TextGrid2 (see Figure 5.2), which has three interval tiers: the manually segmented word tier and two interval tiers generated automatically by AuToBI, one for the pitch accents and the second for the boundary tones.

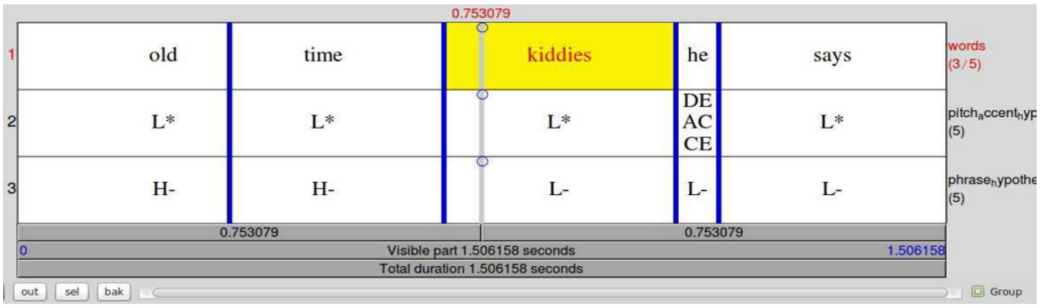


Figure 5.2: AuToBI_v1.0 output: TextGrid2.

AuToBI Output Adaptation Stage (S4)

In spite of the fact that AuToBI meant a great step forward in the systematization of prosodic labeling, it has some major constraints, as has already been mentioned above. Consequently, the information from AuToBI needs to be manipulated to meet our requirements for PPh within the information structure framework. For this purpose, a reduced inventory of ToBI labels was established, as already explained in Chapter 4. One main pitch accent (PA) and boundary tone is labeled in each PPh. Furthermore, while in the standard ToBI convention four tiers of data are foreseen, namely a tone tier, an orthographic tier, a break tier and a miscellaneous tier, only the tone tier is being used.

The automatic adaptation stage is envisaged as a loop of three steps over all sentences of the annotated corpus. The steps are: (1) Initial step, (2) Reduction step, and (3) Pre-revision step.

1. Initial step. This step consists in the assignment of the TextGrid output from AuToBI to the sentence annotated in terms of theme–rheme. The result is a txt file that contains the following fields:

   - Id number of sentence
   - Chain of words
   - Communicative label
   - Chain of ToBI labels

2. Reduction step. The greatest part of the prosodic analysis is carried out during this step of the process. The strings of patterns from Step 2 are envisaged from the perspective of the PPh in the pursuit of establishing not only the possible reduction models, but also the communicative and prosodic criteria to segment long utterances into smaller units. As AuToBI does not predict

bitonals, our reduction step seeks to predict possible bitonals. The following automatic processing is performed on each pitch accent plus boundary tone (PABT) sequence:

- Total deletion of deaccented items (D) or word chains with a low BT (DL%). These intonation patterns match deaccented words, which are disregarded.
- Substitution of deaccented items with a high BT (DH%) by a bitonal marker H+. High BTs in general may provide information on adjacent word stresses, which are relevant in the detection of bitonals when they are followed by a main stress. A sequence of various H+ markers is reduced to a single H+ since it belongs to a sequence of deaccented words. Thus, the resulting single H+ matches a main stress and predicts a bitonal PA.
- Word chains labeled as L*L% in a row can be disregarded for the PPh contour definition. Three-word chains with such a label can be reduced to one L*L% label since only one word in such a chain will be more salient within the PPh.
- Initial L*H%L*L% has been reduced to L*+HL%. In this case, a high BT is turned into a bitonal.
- Three-word combinations of L*H% and L*L% are turned to bitonals with either low or high BTs, depending on the pattern chain. For instance, L*H% L*L% L*H% gives H+L*H%.

The results from this label reduction process are saved into a txt file that contains the following fields:

- Id number of sentence
- Chain of words
- Communicative label
- Number of words
- Number of PPhs
- Proposed ToBI label for each PPh

3. Pre-revision step. After obtaining a PPh label, a Praat file needs to be created in order to manually revise all the material that has been automatically generated. Therefore, a TextGrid3 file merges the existing tiers from TextGrid2 plus three more, namely:

- clauses divided into PPh assigned to their corresponding communicative labels;
- same intonational phrases containing the proposed ToBI pattern, and;
- word divisions as in tier 1 for AuToBI input that will serve to place a pitch accent (PA) into the main stressed word within the PPh and BT, to be able to detect PPhs easily.

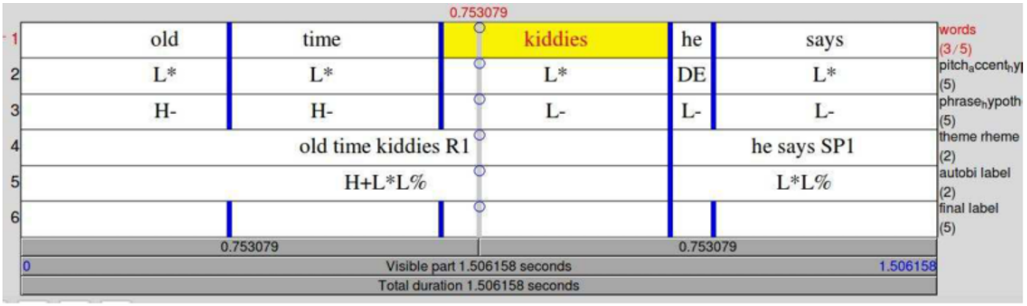


Figure 5.3: Example output: TextGrid3.

Once the Textgrid3 file is generated (see Figure 5.3), the manual process of the validation of the proposed patterns takes place. The manual changes are saved as TextGrid4 (see Figure 5.4).

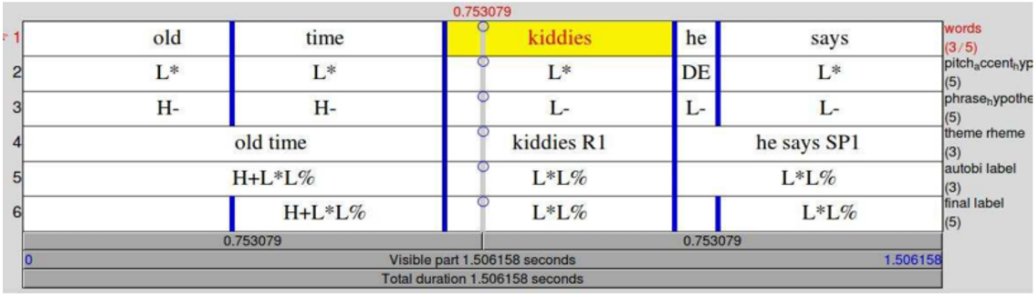


Figure 5.4: Example output: TextGrid4.

Validation Stage (S5)

As mentioned above, the validation stage serves to assess the prosodic patterns obtained during the adaptation stage in order to evaluate the efficiency of our model. For this purpose, the results from the automatic reduction model at the PPh level are compared to a manual annotation. The comparison revealed that the model matches exactly the whole pattern in 58% of the total number of PPhs. This

includes the number of PPh divisions and the exact ToBI pattern assigned. There is a 18% of partially matched patterns (whose match corresponds in all cases to the BT). The remaining 24% of PPh do not match with the manual annotation.

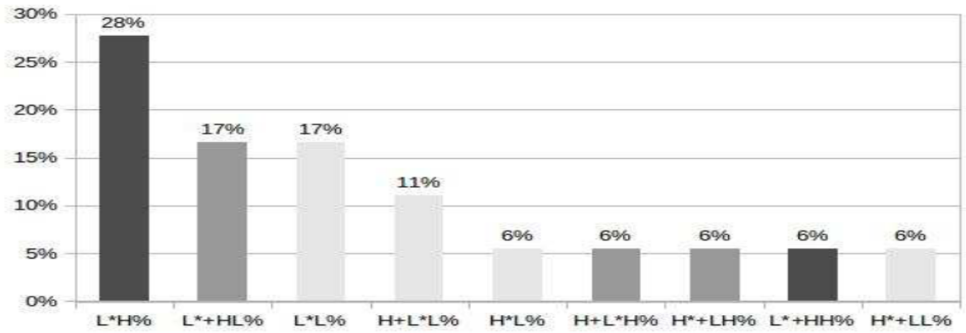


Figure 5.5: Matching patterns from AuToBI_v1.0.

The obtained prosodic patterns are compared with the sentential theme–rheme structures. Figure 5.5 shows that themes tend to contain a rising intonation pattern as he claims, given that L*H%, L*+H H%, H*+L H% and H+L* H% (highlighted in dark gray) have a final rising intonation and L*+H L% contains a rising PA. These patterns add up to 63%, which proves that our model represents the general characterization made in theoretical approaches on this topic. However, this reduction model using AuToBI is still not optimal to establish a fully automatic correspondence between thematicity and ToBI labels as it involves a considerable amount of manual revision at stage (4) and yields a 58% of accuracy when compared to manual annotation.

5.1.2. Experiments on AuToBI v1.5

This subsection describes the revision of the latest version of AuToBI (version 1.5). This version fixes some of the pitfalls in the first version, such as the word-by-word labeling. Nevertheless, some adaptation had to be made on the output. In what follows, a description of the AuToBI raw output on a sample from our corpus is presented, then, I explain how this output is adapted to match ToBI patterns at the PPh and, finally, results of the correlation analysis of these intonation patterns and thematicity are introduced.

AuToBI Output and Post-processing

Even though version 1.5 of AuToBI provides a better output than version 1.0, a post-processing is also needed so that redundant labels are deleted to address prominence and phrasing at the PPh level as established in our methodology (see Chapter 4). Figure 5.6 shows a sample labeling with AuToBI, where three tiers are displayed: (1) the words tier (tier number 2) shows the required word alignment input for AuToBI to run; (2) the tones tier and (3) the breaks tier.

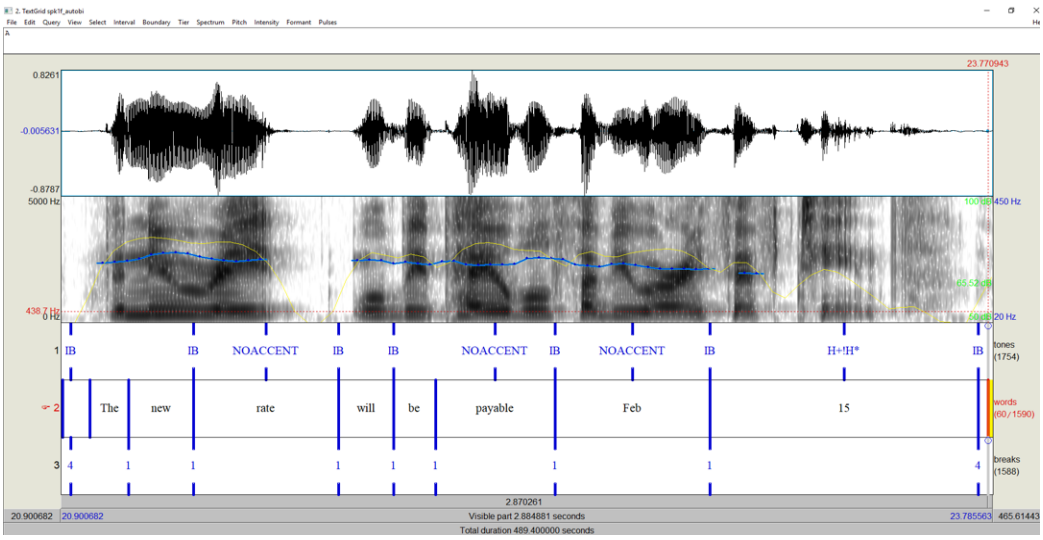


Figure 5.6: Example of AuToBI_v1.5 output.

Event though AuToBI's output in this version does not assign any boundary tone[2], the analysis is conducted to observe the level of coincidence in break assignment and pitch accent patterns with the rising-falling hypothesis corresponding to theme and rheme spans respectively. To this end, AuToBI's output has been post-processed to match intervals in terms of intonational phrases (breaks of type 4) and, at the tones tier, words labeled as "no accent" and "intonational boundary" have been deleted. Consequently, in this analysis, the intonation contour is derived only from ToBI labels for pitch accents.

A conversion algorithm scripted in Praat converts type 4 break points from AuToBI to end boundaries of intonational phrase intervals and writes in each interval the ToBI labels at the tones tier contained within the intonational phrase

[2]Instead the label "intonational_boundary", (abbreviated as 'IB' in Figure 5.6) is inserted in the tone tier.

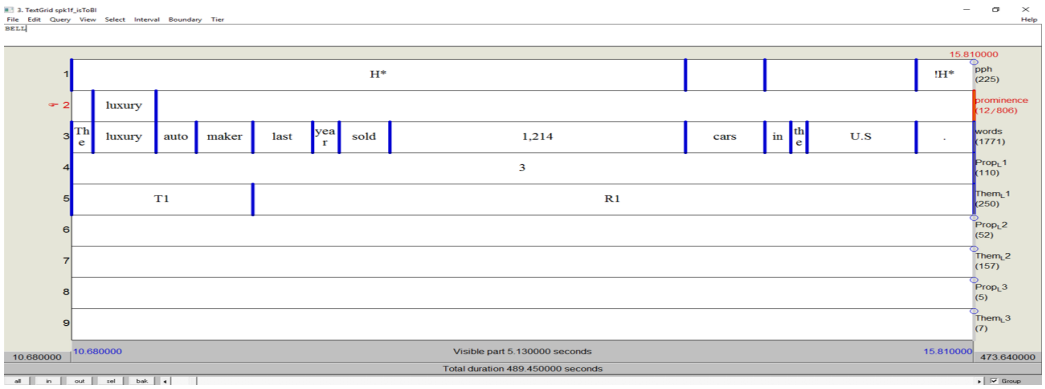


Figure 5.7: Post-processed AuToBI_v1.5 output.

time frame. Words that contain ToBI labels are extracted to a separate tier as prominent words within that intonational phrase. Then, the post-processed AuToBI is merged to the TextGrid with the annotation of thematicity. An example of the resulting TextGrid after post-processing and merging is shown in Figure 5.7. This input is used to compute the coincidence level between intonation and information structure, which is reported in the following section.

Correspondence between Thematicity and AuToBI Labels

The correspondence between thematicity and AuToBI labels is computed in two steps: (1) boundary coincidence and (2) tone pattern characterization of matching spans. The percentage of boundary coincidence between intervals (intonational phrases and thematicity spans) is computed relatively to the total number of thematicity spans (i.e., in the experiments, 504 spans counting all levels of embeddedness). A time margin for boundary coincidence is set at 0.1 seconds to allow partial matches for boundaries, where word boundaries were not exactly used to signal a break point.

Results reported in Table 5.1 show that the level of boundary coincidence relative to thematicity spans is low for the samples of all speakers (an average of 18%); and the highest score does not reach a 30% level of coincidence (i.e, spk3f achieves 27%) for global matches. This result is even lower in tests on perfect boundary matches, which barely attains 20% of the total number of thematicity spans.

Afterwards, ToBI labels are extracted for coinciding spans. Each thematicity span regardless the level of embeddedness is computed. Total matches refer to the relative number in each category of matching spans in the previous experiment.

Table 5.1: AuToBI characteristic patterns for thematicity spans.

Intonation	ToBI pattern	Theme	Rheme	Specifier
Rising	L+H*	**0.15**	0.11	**0.12**
	L*+H	**0.08**		**0.19**
	!H*H*	**0.11**	0.04	0.09
Falling	H*!H*-		**0.17**	
	H*L*-		**0.02**	
	H+!H*	0.17	**0.13**	0.10
Flat	L*L*-			**0**
	H*H*-	0.14	0.20	**0.24**
	H* or L* or !H*	0.29	0.25	0.11
	Total Match	**0.34**	**0.32**	**0.64**

When scores from boundary coincidence and ToBI characterization are combined, results are suboptimal: from the whole dataset, only 5% of thematicity spans coincide with previous theoretical studies on the IS-prosody interface, especially with those that propose a simple theme-rheme division. Results align with the previous analysis when manual ToBI labeling is used (presented in Chapter 6), which suggests that a third span, the specifier, is intonationally distinct from theme and rheme and provides a more fine-grained division of thematicity together with the concept of embeddedness.

A more detailed characterization of both embeddedness and speaker characteristics using the proposed automatic ToBI labeling was not carried out due to the scarce quantity of matches. Taking into account that most of the working corpus is characterized by medium sized sentences (in average fifteen words) containing in 95% of the cases a theme and a rheme at L1, and given that a variety of dialects are included in the voice samples, these results from automatic ToBI labeling do not meet a reasonable level of acceptance to proceed with further testing. Consequently, results of this experiment led us to conclude that automatic labeling of AuToBI is not suitable for the object of the present dissertation.

5.2. Development of Tools for Speech Prosody Annotation

This section presents the implementation carried out based on Praat, thanks to the joint work within the TALN research group, that has lead to the open-source

web platform Praat on the Web[3]. The initial objective was to deploy a rule-based system for the segmentation of speech into prosodic phrases using acoustic parameters. The main motivation was to prepare the grounds for automatic extraction of acoustic parameters in order to automatically create datasets for the exploration of the information structure–prosody interface. The lack of available tools that could comply with our specific requirements led to the implementation of the automatic prosody tagger. The design of the automatic prosody tagger is thought to be versatile enough so that standard processing functions under Praat are scripted in such a way that they can be easily accessed and modified to facilitate adaptation to other tasks, reproducibility of experiments and development of further functionalities. The prosody tagger serves as a benchmark for further exploration of how to segment into other (bigger and smaller) prosodic units. Moreover, researchers dealing with areas related to prosody analysis and generation could further expand the hypotheses presented in this thesis and also explore other interfaces of prosody. For now, Praat on the Web is a demonstration platform where users can upload their speech files to test the segmentation capabilities of the modular prosody tagger. The platform is based on an extension of Praat for feature annotation, which is presented in Section 5.2.1. Then, the automatic prosody tagger is presented in Section 5.2.2.

5.2.1. Praat on the Web: an Upgrade for Feature Annotation

Automatic annotation of speech often involves dealing with linguistic and acoustic information that needs to be conveniently organized at different levels of segmentation (i.e., phonemes, syllables, words, phrases, sentences, etc.). Even though laboratory experiments on speech are controlled to a certain extent (e.g., minimal word pairs, short sentences, read speech) and are usually annotated manually, the increasing trend to analyze spontaneous speech, especially in human-machine interaction, requires tools in order to facilitate semi-automatic annotation tasks with a compact visualization for manual revision, presentation of results and versatile scripting capabilities.

The Praat software (Boersma, 2001) is one of the most widely used open-source tools for audio signal processing and annotation in the speech community. Praat has a dedicated text format called *TextGrid*, where stackable lines, called *tiers*, are mapped to the whole time-stamp of the associated sound file. Accordingly, tiers account for the temporal nature of speech and take one compulsory parameter: the time-stamp of the *segments*, which are the smallest unit in a TextGrid. A time-stamp can be of two kinds: an interval (specifying the beginning and end

[3]http://kristina.taln.upf.edu/praatweb/

time of each segment) or a point in time. These time-stamps form a sequence that is encoded in *tiers*. Once (interval or point) segments are marked, they can take an optional string parameter, called *label*.

As already mentioned, the Praat on the Web tool presented in this section aims to upgrade Praat in accordance with state-of-the-art NLP annotation interfaces, for instance, SEMAFOR[4] (Tsatsaronis et al., 2012), Brat[5] (Stenetorp et al., 2012), and GATE[6] (Cunningham et al., 2011). Such an upgrade is instrumental for prosody studies, among others, which study prosody in connection to various interfaces or as a combination of features (not only acoustic, but also linguistic) and therefore benefit greatly from a versatile semi-automatic approach to annotation and a compact visualization of those features.

Praat on the Web involves three main improvements over Praat: (i) a multi-dimensional feature vector within segment labels (see Figure 5.8 for illustration), (ii) a web-based implementation, and (iii) an operational interface for modular script composition exemplified as a prosody tagger. Given that many Praat scripts are freely available and shared in the speech community for different specialized tasks, one of the advantages of modular scripting within the same platform is keeping a library of scripts for easy replacement of independent subtasks within a larger pipeline. The dynamic configuration approach presented in this section, thus, promotes tests on how different configurations affect the final output of the architecture, and positively impacts reproducibility of experiments in a user-friendly web environment.

Praat on the Web is based on an extension of Praat for feature annotation[7] (also available for local use), and is compatible with the original Praat format as a web application.[8] Source code as well as a tutorial are available in the TALN's repository[9] and distributed under a GNU General Public Licence[10]. In what follows, I elaborate on the improvements introduced in Praat on the Web compared to standard Praat.

[4]http://www.cs.cmu.edu/ ark/SEMAFOR/
[5]http://brat.nlplab.org/
[6]https://gate.ac.uk/
[7]implemented on Praat v.6.0.11
[8]http://kristina.taln.upf.edu/praatweb/
[9]https://github.com/TalnUPF
[10]http://www.gnu.org/licenses/

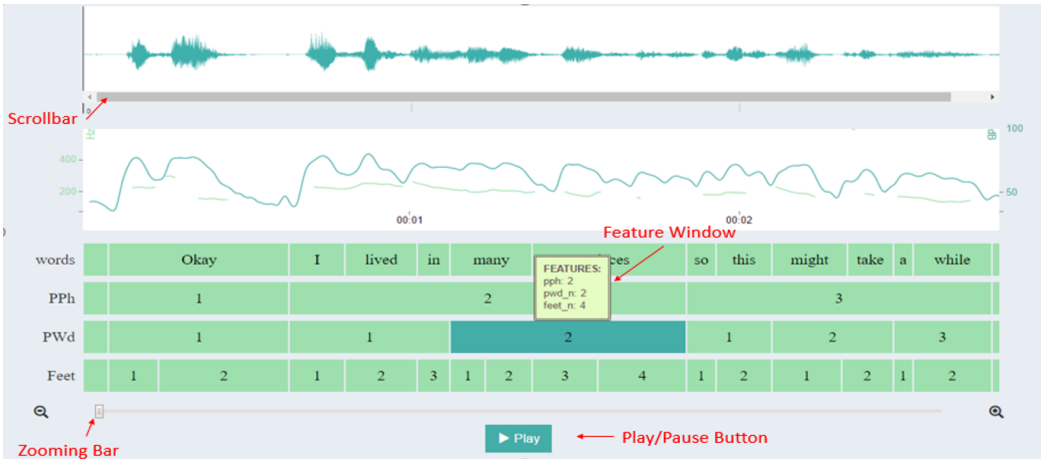


Figure 5.8: Praat on the Web: enhanced visualization interface.

Annotating in parallel tiers versus using features

Annotations in tiers are convenient for studying nested elements in the speech signal. For example, Selkirk (1984) proposes a hierarchical structure of intonation where smaller units (e.g., prosodic feet) are embedded into larger ones. However, if each layer needs to be annotated in stacked tiers with cloned times, a long collection of repeated tiers for each new layer information blurs visual presentation and makes manual revision tasks harder.

The main menu in Praat on the Web includes a first demo (accessible by clicking on the button "Enter Demo 1"), where the user can upload their own audio and TextGrid files for visualization and playback. Sample files with feature annotations, which can serve as examples, are also provided in the demo. Waveform, fundamental frequency (F0) and intensity curves are displayed on the screen together with the annotated tiers. There are some practical differences with respect to the standard Praat, which are summarized in Table 5.2. Whereas standard Praat uses keyboard commands to perform actions during annotation such as zooming and playback, Praat on the Web has dedicated buttons for these actions, as illustrated in Figure 5.8.

Further demonstration of visualization capabilities using automatic scripts for merging tiers and splitting features (Demos 3 and 4 respectively) is also available in the online demo webpage. Users can upload their own cloned TextGrids entering Demo 3 and click on the 'run' button to automatically annotate selected cloned tiers as features. In Demo 4, this action is reversed, i.e., feature vectors are converted to cloned tiers. All TextGrids generated in Praat on the Web are

Action	Standard Praat	Praat on Web
Zooming	keyboard shortcuts (ctrl+i/o/n)	sliding bar signaled with amplifying glass symbol
Audio playback	shift button or segment + time bar click	play/pause button or segment + waveform click
Scroll left/right	scrollbar below TextGrid	scrollbar below waveform

Table 5.2: Actions in standard Praat and Praat on Web.

displayed in the browser and can also be downloaded for local use clicking on the "Download" button.

Dynamic Scripting Configuration

Entering Demo 2 through the main menu of Praat on the Web, an example of dynamic scripting composition can be run on available samples or uploaded files. The configuration of the automatic prosody tagger appears in the right part of the screen (see Figures 5.9 and 5.10). Further information on the prosody tagger methodology, technical specifications and evaluation is provided in Section 5.2.2.

The pipeline varies depending on the selected configuration. The prosody tagger is made up of a total of eight modules, three of which (from Module 1 to 3) are common for the two possible configurations:

1. Word segments (see Figure 5.9): when clicking on this button, six modules will appear in the "Selected modules" box. Modules 5 and 6 predict boundaries and prominence respectively on both acoustic information annotated in Modules 1 to 3 and word segments exported by Module 4. A TextGrid with the word alignment needs to be provided to run this configuration.

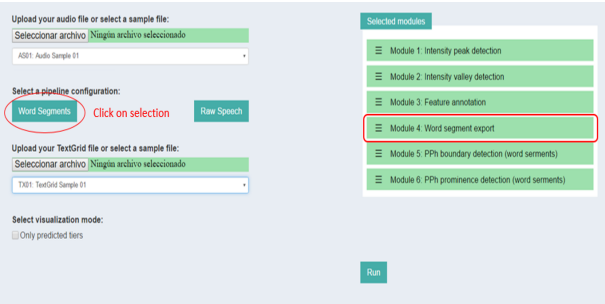


Figure 5.9: Configuration with word segments.

2. Raw speech (see Figure 5.10): when clicking on this button, five modules will appear in the "Selected modules" box. Prediction is performed on

acoustic information and, thus, Module 4 is not in the pipeline and alternative Modules 5 and 6 are chosen for this pipeline.

Figure 5.10: Configuration for raw speech.

The users can select in the web interface the output of the prosody tagger by ticking the option "only predicted tiers" displayed at the bottom left side of the screen. If that option is not ticked, all tiers generated by each module are shown. The output of the tagger (including annotated features of each segment) is displayed on the screen in the browser; it can also be downloaded in TextGrid format for local use.

A further add-on of Praat on the Web is that includes a centralized repository of scripts and data. The action of selecting modules for the sample prosody tagger has been scripted in this demonstration to be automatically done; the web interface allows moving around modules to prove that modules are also manually interchangeable.

5.2.2. The Automatic Prosody Tagger

Communicatively-oriented prosody has been proved to be central for advanced speech technologies. It is decisive in structuring the message, stressing parts of the message that the interlocutor considers important, and revealing information about the interlocutor's attitude and affection state (Nooteboom, 1997; Wennerstrom, 2001). However, despite the advances of theoretical studies in the information structure–prosody interface, so far, no sufficiently large, annotated prosody material has been created to support empirical studies and drive the research on empirical techniques for analysis and generation of prosodic cues based on communicative approaches, especially for application in human-computer interaction technologies.

Common annotation conventions, such as the ToBI convention, provide a descriptive framework of intonation contours and phrasing based upon labels that are language-dependent and rather subjective, which makes it difficult to reach a satisfactory inter-annotator agreement for creating gold standard annotations to train and evaluate algorithms.

It is, therefore, not surprising that empirical research on the information structure–prosody interface is still based upon rather small laboratory experiments in the best of the scenarios. A further consequence of the lack of sound universal prosody annotation conventions is that current methodologies applied to speech prosody segmentation are still based upon textual and linguistic units (usually words or syntax) rather than on acoustic and phonological units (prosodic phrases and prosodic words). These limitations become an insurmountable barrier for technologies that aim at grasping prosodic cues, especially in spontaneous speech, where many complex prosodic, linguistic and affective phenomena occur (hesitations, incoherent discourse structure, false starts, continuation rising tunes for holding the floor, expression of emotions, speech acts, prosodic disambiguation, etc.).

The inherent peculiarities of oral language cannot be dealt with using strategies that belong to written language. For instance, sentences with false starts including a filled pause (e.g., *They've never ... mmm well, my brother's been to Barcelona*). To overcome the limitations of the current annotation practice and advance in the derivation of more meaningful communicative units from speech as well as in the generation of more natural synthesized speech, the following issues must be tackled:

- a parametric language-independent annotation schema of prosody at the acoustic level that can be used by computational models for automatic segmentation and prominence detection;
- prosody taggers and acoustic feature extractors that distill acoustic features from raw speech signals.

In what follows, both tasks are addressed as an implementation experiment of a modular tool that segments spontaneous speech using a parametric approach and a set of rules. The prosody tagger is deployed in the extended version of the Praat for feature annotation previously explained in Section 5.2.1. Such a feature annotation functionality contributes to the independent modular structure and also helps visualization and manual revision of the output within the same Praat environment. The inter-annotator agreement figures and tagger performance compared to a baseline using only F0 cues show that our work is a relevant contribution

to the state of the art in the field of speech prosody processing.

The rest of the section is structured as follows. Firstly, the architecture and technical description of the prosody tagger are described. Then, the inter-annotator agreement and evaluation of the performance of the prosody tagger is presented. Finally, the contributions of the prosody tagger are summarized.

Prosody Tagger Implementation

The automatic prosody tagger is available as a web service running as part of Praat on the Web[11]. Any speech sample in *wav* format and associated *TextGrid* with word segments can be uploaded for processing. The segmentation into PPhs and prominence within these prosodic units is displayed on the screen and also available for download in TextGrid format for local use. All scripts and the extended Praat version for feature annotation are available under a Creative Common's license[12].

The architecture of the prosody tagger has been designed as a modular platform such that its optimization and further development can be attained focusing on specific intermediate steps within the whole pipeline. Acoustic information extracted from different modules is annotated as a feature vector in each segment, including computed z-scores within different prosodic units (so far, from levels 1 to 3). Acoustic parameters include so far, but are not limited to, F0, intensity and duration elements, as Praat allows extraction of a wider range of acoustic parameters (such as jitter, shimmer and pulses, among others).

Figure 5.11 sketches the modular architecture of the prosody tagger with two possible configurations, as exemplified in Praat on the Web: (i) Default 1: using only raw audio (as wav file), and (ii) Default 2: using both raw audio (wav file) and importing external word segmentation (in TextGrid format), which must be uploaded by the user. For the Default 2 configuration presented in this study, we have used for word segmentation the proprietary Automatic Speech Recognition system *Scribe*[13] by Vocapia Research[14]. The output of Scribe is converted from xml into TextGrid format. In what follows, a description of each module's functionality is outlined and annotated acoustic features are specified at each stage.

[11] http://kristina.taln.upf.edu/praatweb/
[12] https://github.com/TalnUPF/
[13] https://scribe.vocapia.com/; Scribe is currently run as a beta version.
[14] http://www.vocapia.com/

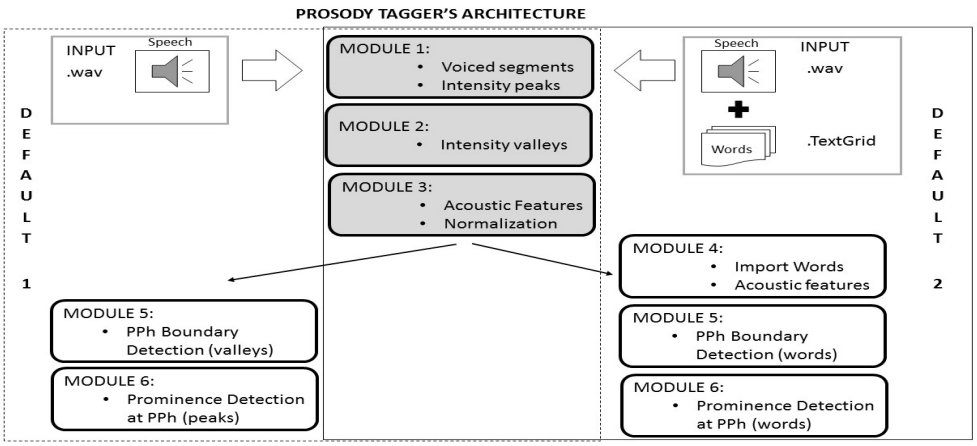


Figure 5.11: Prosody tagger architecture.

Module 1 uses the wav file and creates a TextGrid using the built-in function in Praat *To TextGrid (silences)*, which automatically detects unvoiced and voiced segments as intervals. Then, a pitch object and an intensity object are extracted from the sound file. The function *To IntensityTier (peaks)* is performed on the intensity object to select salient peaks. The F0 information is extracted at standard frame rates from the pitch object to associate extracted intensity peaks to the ones that involve F0; the distance between these peak candidates is also considered for syllable nuclei detection. A point tier is created and points matching the combination of intensity, F0 and time distance within each voiced segment are annotated. As features, absolute intensity, F0 and the associated voiced interval are stored in each point segment.

Module 2 makes use of the intensity object created in Module 1 to extract intensity valleys using the Praat function *To IntensityTier (valleys)*. Standard intensity frames are selected if their intensity z-score (relative to L1) is lower than 0. Then, the lowest values in intensity relative to each voiced fragment (L2) are labeled in a new point tier taking into account the distance between them. Annotated features from this module are: intensity z-score relative to the whole sound (L1) and intensity z-score relative to the associated voiced segment (L2) at each valley point.

Module 3 extracts acoustic values, computes z-scores at available levels, and annotates results as features in each segment. At L1, mean and standard deviation of intensity and F0, together with duration for the whole file, are annotated. These values serve for calculation of z-scores at lower levels in the hierarchy. At L2,

annotated features include both absolute values for F0, intensity and duration for further calculation of z-scores in peak and valley tiers (created in Module 1 and 2 respectively) and z-scores derived from L1 values. In the peak and valley tiers, the distance to the previous point is also annotated as a feature. For the first point in the tier, the distance to the boundary of its associated voiced segment is specified as reference.

When a TextGrid with the word segmentation is available (upon selection of the Default 2 configuration), Module 4 exports this tier and annotates features at each marked interval. Consequently, prominence predicted in Module 6 outputs prominent words, given that word alignment is provided by the user. Extracted acoustic parameters and annotated features in this module include: (i) z-scores relative to their associated L2 voiced interval (the z-score values for intensity and F0 are extracted and annotated as features for each word segment obtained by Module 1); (ii) time landmarks, i.e., time of minimum value of intensity and maximum F0 within each word; (iii) duration: absolute duration of the word, and relative duration to the corresponding voiced segment and to the whole sample.

Module 5 uses voiced segments and valleys to predict PPh boundaries. They are derived from the information extracted in the L2 voiced/silence segments detected by Module 1 and from the valleys marked in Module 2. In each L2 voiced segment, the smallest z-score values of intensity are tagged taking into account the distance of these valleys to the closest peaks. If the distance of one of the closest peaks is greater than or equal to 0.2 seconds, the z-score is among the minimum in the range, and F0 value is *undefined*, then a PPh boundary is marked.

Finally, Module 6 performs prominence detection on each PPh predicted in the previous module. If no word alignment is available, only syllable peaks predicted in Module 1 are used. Consequently, this module outputs prominent points that correspond to peak points in configuration 1 and prominent word intervals in configuration 2. For calculation of prominence, a combination of F0, intensity and duration cues are taken into account as described in Chapter 4. Figure 5.12 displays all tiers created by each module as described above for computation and the final output with the tagging of PPh boundaries and prominent words after the whole pipeline has been executed running a default 2 configuration.

Evaluation

A total of five different spontaneous speech samples have been used in the evaluation, both for inter-annotator agreement and the performance of the prosody

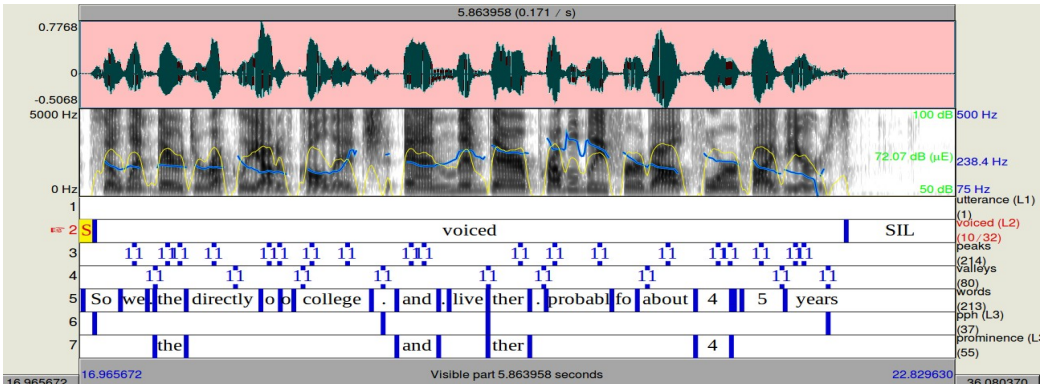


Figure 5.12: Output tiers by each module from the prosody tagger.

tagger: three dialogs in Spanish and two monologues in American English. Table 5.3 shows the specific information details for each sample.

Table 5.3: Corpus used in the evaluation of the automatic prosody tagger.

Filename	Format	Length	
		Seconds	Words
es01mm	dialog	36	196
es02mm	dialog	28	150
es03fm	dialog	152	545
en04m	monologue	70	213
en05f	monologue	30	282
	TOTAL	**316**	**1386**

Dialogs in Spanish are set in a medical context; a male doctor is involved in all of them talking to a patient. Gender is represented in all file names with the convention “f” and “m” for female and male respectively. Files “es01mm” and “es02mm” include the same speakers in the same conversational context, where a patient complains to the doctor. In “es01mm”, the doctor shows a negative response, while in “es02mm”, he acts in a comprehensive and pro-active way. Monologues in English are biographical introductions of the speakers (birthplace, family, recent activities, etc.).

Two expert annotators, proficient in both English and Spanish, have independently labeled both speech samples following the guidelines outlined above in Chapter 4. Cohen’s kappa (Cohen, 1960) has been calculated for inter-annotator agreement for each prominence and boundary labeling task. Evaluation is performed on the Default 2 configuration using word segmentation to facilitate the

computation and objectiveness of the validation process. A baseline pipeline using only duration and F0 parameters for the same task has been implemented. Inter-annotator kappa results follows and the accuracy of the tagger, precision and recall compared to the baseline is reported thereafter.

Inter-annotator agreement Table 5.4 provides kappa values for PPh boundary and prominence labeling of our corpus. If annotators label words that are part of the same prosodic word (e.g, they coincide in the final initial word boundary or are separated by a function word, which is usually unstressed), this counts as a partial match for the kappa computation. In order to count matches automatically under Praat, annotators were asked to insert interval boundaries duplicating the word boundaries that are automatically marked, so that it is possible to compare boundary times for the computation of matches.

Table 5.4: Inter-annotator agreement: Cohen's kappa.

Filename	Prominence	Boundary
es01mm	0.55	**0.98**
es02mm	**0.63**	**0.72**
es03fm	0.51	**0.78**
en04m	**0.72**	**0.93**
en05f	**0.69**	**0.70**

A kappa within the range of 0.6-0.8 (within a scale between 0 and 1) is considered *satisfactory*, and above 0.8 *perfect* (Cohen, 1960). In Table 5.4, kappa values that are in line with these thresholds are highlighted in bold for each task (i.e., prominence and boundary labeling within PPh level). Results prove that agreement ranges from 0.51 and 0.98. A higher agreement is observed in the boundary labeling task for all voice samples. No significant differences are observed between English and Spanish samples in boundary detection. However, in prominence labeling, two Spanish samples (files "es01mm" and "es03fm") only reach a *moderate* agreement of 0.55 and 0.51 respectively and, in the overall picture, kappa values for prominence in Spanish are lower than those for English, which might be due to the dialog format of the samples with shorter interventions, affective states displayed by participants (perceived emotional behavior conveyed by prosody) and quick turn movements between speakers.

It cannot be inferred that prominence annotation could be language-dependent as such, as the corpus used for the evaluation is simply too small for such conclusions. Nevertheless, further research could exploit these techniques, or even

a combination of semi-automatic annotation using the prosody tagger presented in this section, to explore how linguistic parameters such as the discourse type (dialog in Spanish versus monologue in English), register, gender or speaker idiosyncrasies may affect inter-annotator agreement and tagger's performance in this respect.

Automatic prosody labeling performance In order to evaluate the performance of the automatic prosody tagger, full matches are considered when the output of the tagger matches either one or both annotators. For prominence labeling only, partial matches, i.e., words that coincide in one interval boundary or belong to the same prosodic word are also considered as a match. Boundaries that match with a time margin of ± 0.25 seconds are considered to be partial matches. For PPh boundaries, a match is counted if the automatic tool labels a boundary which has only been labeled by one annotator.

Table 5.5: Automatic prosody tagger evaluation.

	Accuracy		Precision		Recall		F-Measure	
	P	B	P	B	P	B	P	B
baseline (F)	0.83	0.89	0.49	0.88	0.22	0.28	0.30	0.42
tagger (F)	**0.84**	0.88	**0.52**	0.58	**0.32**	**0.43**	**0.36**	**0.55**
baseline (F&P)	0.90	0.90	0.84	0.88	0.37	0.28	0.51	0.42
tagger (F&P)	**0.91**	0.89	0.80	0.63	**0.49**	**0.49**	**0.61**	**0.55**

Table 5.5 presents the accuracy, precision, recall and f-measure scores for full matches (F) and full and partial matches (F&P), both for the baseline and our tagger. Results show that the prosody tagger performs at accuracy rates higher than 0.84 in both prominence (P) and boundary (B) detection tasks. The baseline achieves higher precision figures (especially in boundary detection) than our tagger. A closer look at the output reveals that the baseline marked only those boundaries that included a clear pause, i.e., "safe" candidates. In contrast, the tagger marked not only those clear pauses, but also more subtle boundaries that involved an intensity decrease and not necessarily a pause. On the other hand, the tagger reaches considerably higher recall figures than the baseline for both prominence and boundary detection tasks. The f-measure figures show that overall, the tagger performs better. Still, since our methodology is based upon the deviation of normalized values, neutral speech might pose a problem when trying to tag both prominence and boundaries, as there is a tendency towards less variable prosodic cues in this register. Further empirical studies using a semi-automatic approach

and optimization of the tagger are needed to have a deeper insight into this issue.

5.3. Discussion

This chapter elaborates on the hypothesis that automatic labeling of prosody facilitates the analysis of the information structure–prosody correspondence. Due to the need of exploring large amounts of data to obtain good results in training classification algorithms, the exploration of an existing automatic tool for the annotation of ToBI, AuToBI, has been carried out at different stages of development. Two versions of AuToBI were tested. However, results are considered suboptimal for the working corpus of read speech compiled for this dissertation.

As reported in Section 5.2, the implementation experiment of the automatic prosody tagger proves that an acoustic representation of prosody based on rules serves to segment spontaneous speech in two languages (English and Spanish). Moreover, the inter-annotator agreement figures show that the annotation schema for the annotation of prominence and phrasing introduced in Chapter 4 does not depend on potentially subjective criteria of the individual annotators. Besides, the recall results prove that the automatic prosody tagger is flexible enough to support language independent speech signal analysis and detection of prominence and boundaries at the PPh level using a combination of acoustic features, rather than merely F0 contours, as previous empirical and theoretical studies claimed. Improved recall scores also indicate that the number of true positives from the total number of words which actually belong to the positive class, i.e., labeled as positive by the manual annotators, is higher than the baseline.

The implementation of the prosody tagger may be extended for other applications or further smaller prosodic unit detection (such as prosodic words) due to its modular architecture. Moreover, the extended Praat functionality for feature annotation running on the web platform Praat on the Web provides easy access and manual revision of the output of the prosody tagger. Furthermore, Praat on the Web aims to meet the increasingly demanding requirements in the field of speech technologies. User-friendly semi-automatic annotation tools within one versatile common platform are key to make steady progress in the study of complex events, like prosody, over large amounts of data. Praat on the Web shows several advantages over the standard Praat in that it offers:

- intuitive visualization of segment annotations using features displayed in a dedicated window;
- easy modularity of computational tasks within the same Praat platform;

- ready-to-use web environment with no pre-installation requirements for presentation of results.

The two first characteristics are achieved including the functionality for feature annotation. Thanks to this added functionality, the smallest unit in a Praat TextGrid is no longer an opaque string label, but a well-structured linguistic unit containing a *head*, a *feature name* and a *feature value*.

At the time of writing this dissertation, Praat on the Web runs with sample or uploaded files for visualization, playback and automatic prediction of PPh boundaries and prominence. In a future release, user account management could be introduced for researchers to upload their scripts and create their own pipeline configurations. So far, the web interface is well-suited for annotation demos and teaching purposes; a further extension with online edition of manual annotations would also be an asset.

All in all, the presented methodology and implementation serves as a platform upon which further research lines and experiments can be run to increase the knowledge in the area of speech prosody and test advanced implementations for the automatic annotation of speech prosody.

Chapter 6

EXPERIMENTS ON THE INFORMATION STRUCTURE–PROSODY INTERFACE

> ”In so far as such a theory is empirically correct, it will also tell us what empirical facts it should be possible to observe in a given set of circumstances.”
>
> — Talcott Parsons

This chapter unfolds around experiments that provide empirical evidence to support theoretical constructs on the information structure–prosody interface. The first goal is to contribute to prove and extend previous hypotheses on the information structure–prosody correspondence. The findings will ultimately serve to implement a versatile prosody module for generation of communicative synthesized speech. Thus, the objective is to analyze the information structure–prosody correspondence from different angles: from the theoretical perspective, using a corpus-based approach; and, from the practical point of view, in an implementation setting within a CTS application.

The corpus-based analysis of the information structure–prosody interface is presented as a set of questions on the topic. Statistical analysis and classification experiments are used as a means to explore the answers to the proposed questions. The first section explores the corpus of study itself. It is well documented in the literature that different dialects, age and gender (among other factors) often undergo specific (and, thus, differentiating) intonation cues across speakers; see, e.g., (Rose, 2002) for a complete review on how phonetic differences serve

to identify speakers. Consequently, it seems reasonable to commence the chapter on empirical experiments by validating the suitability of the working corpus before any other experiments are carried out. Hence, Section 6.1 answers the first question: *What common prosodic features does our corpus have?* It is explored to what extent the proposed methodology for prosody representation accounts for common prosodic features across speakers in our corpus.

After this verification, it is justified in Section 6.2 why using tripartite (specifier–theme–rheme) hierarchical thematicity is more convenient than the traditional binary (theme–rheme) flat division. From here, I move on to the non-trivial question at stake in this dissertation: how to bridge the gap between theoretical studies on the information structure–prosody interface and its applications. As the actual implementation of prosodic variations needs to be carried out by means of acoustic parameters, labels for the representation of prosody must be eventually mapped onto the acoustic signal for testing in synthesized speech generation settings. In Section 6.3, experiments are carried out to explore to what extent ToBI labels relate to three acoustic elements.

After all these questions have been solved, Section 6.4 describes a corpus-driven approach to analyze how prosody is related to hierarchical thematicity using ToBI labels and acoustic parameters. A final implementation experiment is carried out in Section 6.5 on thematicity-based prosody enrichment of synthesized speech within a CTS application.

6.1. What Common Prosodic Features does our Corpus Have?

Previous theories on the theme–rheme correspondence with rising–falling patterns presuppose that all speakers signal thematicity and that they all do it using the same intonation patterns. However, I put this assumption to the test in the first place to observe if this is really so in our corpus of read speech. In the compilation of the corpus, one of the requirements was to gather samples from different genders, age ranges and dialectal regions in the USA to observe whether the information structure–prosody correspondence is generalizable across a range of native speakers of American English.

Initially, our representation of prosody does not take into account speaker dependent phonetic variations of prosody. On the one hand, the annotation of ToBI

patterns proposed in this dissertation is meant to represent overall F0 movements without considering, for example, syllabic realization of those F0 movements.[1] On the other hand, acoustic parameters are extracted as mean values within specific spans, so they do not account for idiosyncratic variations as such. But, is this approach enough to eliminate prosodic features that are speaker dependent?

In this section, experiments test to what extent the proposed prosody representations account for common prosodic features across speakers . To this aim, I propose a preliminary statistical experiment on the analysis of variance in Subsection 6.1.1. In Subsection 6.1.2, a comparison is carried out between a detailed ToBI annotation (as specified in the ToBI annotation guidelines (Silverman et al., 2010) and online tutorials) and the reduced set of ToBI labels presented in Chapter 4. Furthermore, a classification experiment is proposed as a way to analyze the combination of acoustic parameters to predict ToBI labels comparing five speakers from different dialectal regions. As a conclusion to this section, a discussion is presented in Subsection 6.1.3.

6.1.1. Statistical Analysis of Prosodic Parameters

A statistical test for the analysis of variance is proposed as an exercise to prove whether the analyzed prosodic parameters are speaker dependent in our corpus and, therefore, not generalizable, or whether they are homogeneous across speakers and thus valid for inclusion in the characterization of the information structure. A one-way ANOVA is conducted to compare acoustic parameters within speakers using all spans in each level of segmentation. Table 6.1 reports the degrees of freedom ('df'), the F value ('F') and the significance value ('Sign.') of this test.

Table 6.1 shows that there are statistically significant differences (highlighted in bold) in all absolute values and in the following normalized values: z_int, z_f0 and maxf0.t. This suggests that prosodic prominence is realized by different ranges of relative intensity and F0 (i.e., z_int and z_F0), and that the maximum F0 is located in different positions (i.e., maxf0.t) when all speech samples from the twelve speakers are considered. The rest of the acoustic parameters did not significantly differ between speakers, in particular, relative values extracted from previous spans (i.e., z_int_p, z_f0_p and z_dur_p).

Post hoc comparisons using the Tukey HSD test indicate that mean scores

[1] L*+H refers to both post nuclear rising accent, e.g., in the word *Mary* with a lexical L* stress on the syllable *Ma-* with a rise in F0 in the post-nuclear syllable *-ry* versus a nuclear rise within the lexically stressed syllable as in *John*, labeled as L+H* in the ToBI convention.

Table 6.1: Results from one-way ANOVA between speakers.

	df	F	Sign.
z_int	F(11,6024)	2.34	**0.007**
z_f0	F(11,5986)	4.67	**0.000**
z_sr	F(11,6024)	0.53	0.883
z_int_p	F(11,4716)	0.61	0.826
z_f0_p	F(11,4678)	1.67	0.073
z_dur_p	F(11,4716)	0.00	1.000
maxf0.t	F(11,5986)	7.49	**0.000**
minInt.t	F(11,5986)	7.49	**0.000**
int(dB)	F(11,6024)	143.62	**0.000**
f0(Hz)	F(11,5986)	3206.84	**0.000**
dur	F(11,6024)	8.48	**0.000**
sr	F(11,6024)	26.31	**0.000**

for z_f0 of spk6m, spk4m and spk2f were significantly different than the rest, as shown in Table 6.2.

Taken together, these results suggest that relative acoustic parameters, in particular those relative to previous spans and relative speech rate, do not differ greatly between speakers in the corpus, whereas absolute acoustic parameters show statistically significant differences between speakers. Normalized scores of maxF0.t also differ considerably in our corpus, which suggests that even though F0 contours are not studied in detail (e.g., considering variations across syllables), the corpus contains significant differences in F0 prominence location. Such evidence supports the argument that F0 prominence may be dependent on linguistic features and speaker choices.

Table 6.2: Post-hoc Tukey test for z_f0.

	Mean	Std
spk2m	-0.01	0.63
spk3f	-0.01	0.55
spk6f	-0.01	0.57
spk1f	0	0.61
spk5f	0.01	0.52
spk5m	0.01	0.61
spk3m	0.02	0.64
spk7f	0.04	0.60
spk1m	0.07	0.67
spk4m	**0.10**	**0.92**
spk2f	**0.12**	**0.75**
spk6m	**0.21**	**1.2**

PostHoc tests also demonstrate that in the corpus, there are three speakers who differ greatly in the range of F0 variations. In any case, the F0 parameter will be included for the study of the thematicity–prosody correspondence, as F0 is the main indicator of intonation. In order to further test if these results are consistent when the number of spans is reduced to only those containing simple sentences (i.e., L1 propositions), a one-way ANOVA between speakers was conducted at the sentence level.

Table 6.3: Results from one-way ANOVA between speakers using sentence spans.

	F(11,1296)	Sign.
z_int	0.39	0.958
z_f0	0.23	0.996
z_sr	0.09	1.00
maxf0.t	5.98	**0.000**
minInt.t	1.69	0.070
int(dB)	159.24	**0.000**
f0(Hz)	4589.68	**0.000**
dur	9.16	**0.000**
sr	26.32	**0.000**

Results from the test using sentence spans are reported in Table 6.3. Absolute values and maxF0.t also show statistically significant differences between speakers. However, it should be noted that z_f0 does not show significant differences at the sentence level. As a conclusion from these experiments, absolute acoustic parameters and time location of the F0 peak (i.e., maxF0.t) cannot be considered common prosodic features in the corpus. Consequently, they will not be included in the parametric characterization of thematicity. On the contrary, normalized values, especially z-scores calculated in relation to the previous span, are suitable for characterizing thematicity. Having shown that acoustic parameters related to F0 tend to be speaker dependent, the next question that arises is whether ToBI, being mainly a representation of F0 contours, contains speaker dependent characteristics. The following section explores this question.

6.1.2. Classification Experiments

Two classification exercises are carried out in this section. The first experiment aims at proving the advantages of a reduced catalog of ToBI labels (as defined in Chapter 4) instead of an annotation including the whole range of possibilities accepted in the ToBI convention. The second experiment analyzes how acoustic parameters relate to ToBI labels using five speakers whose state of origin (i.e.,

New York, Illinois, Texas, Massachusetts and Arizona) belongs to different dialectal regions in the USA.

The following reduced datasets are employed: AL5FD, AL5FTD and ALRD. A combination of linguistic and acoustic features is used as attributes to predict ToBI labels. The class to be predicted is a ToBI label, each instance consists of a word represented by linguistic attributes –including PoS, syntax, thematicity and location in the sentence and thematicity span, and acoustic features –including range, minimum, mean and maximum intensity and F0, and duration of the word. A J48 classifier with 10-fold cross-validation is used in both classification exercises.

Comparison of ToBI Annotation Schemas

One speaker (spk5f) was considered to test whether our annotation schema with a reduced set of ToBI labels is more appropriate than a detailed annotation of the intonation contour that follows the specifications given by the ToBI convention. Thus, the following datasets are compared: AL5FD (with the reduced ToBI label catalog) and AL5FD (with the detailed ToBI annotation). Attributes and instances are the same in both datasets, only the number of classes to be predicted varies from 9 in AL5FD to 28 in AL5FTD. A ZeroR classifier is used as a baseline (BL). The absolute improvement (AbsImp) of a J48 tree classifier with 10-fold cross-validation (J48) over this baseline is used as the assessment metric.

Table 6.4: Comparison of ToBI annotation schemes.

Annotation	Dataset	n.classes	A (BL)	A (J48)	**AbsImp**
ToBI convention	AL5FTD	24	22%	43%	11%
reduced ToBI catalog	AL5FD	9	30%	60%	**30%**

Table 6.4 shows an absolute improvement in accuracy of 30% of the reduced ToBI catalog in comparison to an 11% improvement of the detailed ToBI annotation for spk5f. There is a direct correlation between reducing the number of classes and attaining better results in classification experiments. However, as the absolute improvement over a majority voting strategy (ZeroR classifier) is being used as metric to assess the prediction capability, the probability of classifying classes as the class with the highest number of instances is not taken into account.

Analysis of Acoustic Parameters in ToBI labels

Once the reduced catalog of ToBI labels has been proved to yield a higher absolute improvement on speech samples from one speaker, I set out to analyze if this reduced set of ToBI labels maps onto a specific combination of acoustic parameters. To this aim, individual datasets from five participants are used, namely: AL1FD, AL1MD, AL2MD, AL4MD and AL5FD.

In this experiment, accuracy for predicting ToBI labels from different combinations of linguistic and acoustic features are compared for each dataset. Datasets differ in terms of the associated acoustic parameters (extracted automatically using Praat) for each instance. Linguistic attributes are always included in the prediction. Only variations regarding prosodic elements (intensity, F0 and duration) are being tested. In this exercise, each dataset is considered separately from the rest, that is, the prediction problem reflects to what extent ToBI labels are affected by a combination of three prosodic elements, on the one hand, and whether there are common tendencies among speakers in this respect.

Figure 6.1 contrasts, for each sample, the prediction accuracy of the combination of linguistic attributes with each acoustic element individually against the combination of linguistic attributes with all three acoustic elements together. It is shown that for each individual speaker, the combination of linguistic attributes with three acoustic elements leads to a higher performance than a combination with only one acoustic element. The maximum prediction accuracy achieved using the combination of all linguistic and acoustic attributes is 63%. This modest result suggests that ToBI labels may depend on linguistic levels that are not considered in this prediction experiment.

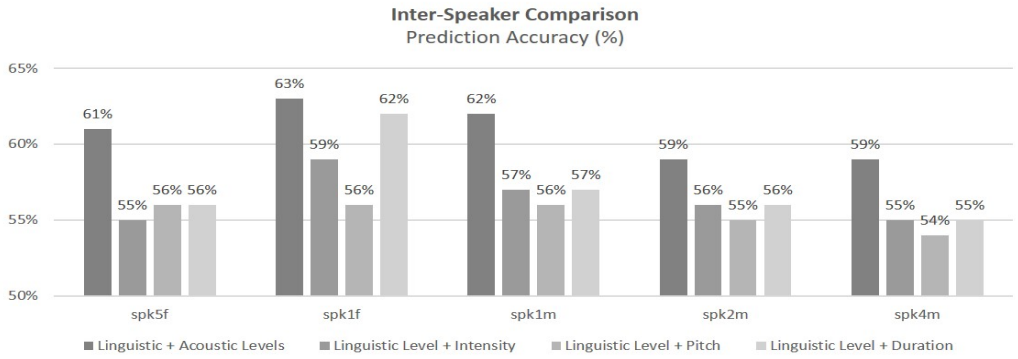


Figure 6.1: Comparison of inter-speaker accuracy.

With respect to the distribution of accuracy values for each acoustic element, in spk1f, for instance, the duration element achieves results (62%) that are close to the combination of all features (63%). In all other speakers, even though results are not exactly the same, each element attains a similar accuracy result with a difference of $\pm 1\%$. Furthermore, for all speakers except spk5f, the lowest accuracy is obtained when the F0 element is used in isolation.

The speaker's choice on making shorter or longer PPh and placing the PA in one word or another may lead to one sentence having two different possible prosodic realizations, as shown in Figure 6.2 for the sentence *Mr. Mayor's hope that references to "press freedom" would survive unamended seems doomed to failure*[2].

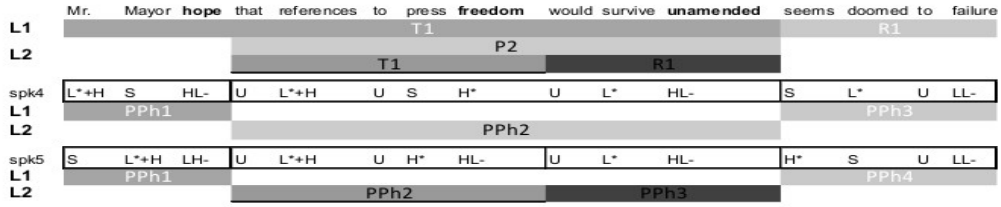


Figure 6.2: Example of hierarchical thematicity and PPh matching.

Figure 6.2 sketches the existence of common prosodic markers that are consistently located on the same words by different speakers, which establishes broad PPh divisions that coincide with different hierarchical levels of the thematicity structure. PPh2 of speaker 4 ('spk4') matches the whole proposition 2 (P2), while speaker 5 ('spk5') splits P2 into two PPh, i.e., PPh2 coincides with T1(P2) and PPh3 with R1(P2). This example shows the importance of annotating thematicity over propositions and including embedded thematicity for finding matching prosodic phrases over those text spans.

6.1.3. Discussion

The statistical analysis presented in this section proves that some speakers show variations in parameters related to F0. Previous phonetic studies (e.g., (Rose, 2002), among others) prove that intonation is idiosyncratic, i.e., speakers have their own characteristic way of using speech prosody. Other studies emphasize

[2]Sentence 57 in our corpus.

the importance of distinguishing between phonetic and phonological transcriptions of prosody to account for differences and commonalities in prosody across speakers, and even across languages; see, e.g., (Hualde, 2000; Hualde and Prieto, 2016).

Despite the fact that the differences between phonetic and phonological transcriptions of prosody is out of the scope of the present dissertation, it is important to mention that from a methodological point of view, the prosody representation chosen to study the information structure–prosody interface needs to be conveniently tested in corpus-driven approaches such that it is possible to generalize results.

6.2. Why Hierarchical Thematicity?

This section provides empirical evidence that supports the hypothesis that hierarchical thematicity is appropriate to study the information structure–prosody interface. Subsection 6.2.1 proves the advantage of the tripartite hierarchical versus a binary flat thematicity. Then, I demonstrate to what extent hierarchical thematicity (together with other linguistic features) contributes to the prediction of ToBI labels in Subsection 6.2.2. A final experiment is devised in Subsection 6.2.3 to explore the bidirectional correspondence between acoustic parameters and thematicity labels in our corpus. Subsection 6.2.4 closes the section with a discussion on the results.

6.2.1. Binary Flat versus Tripartite Hierarchical Thematicity

The first argument that supports the hypothesis that a tripartite hierarchical thematicity is more appropriate than a binary flat thematicity is the possibility of covering a wider spectrum of sentence complexity using a formal representation of communicative structure, as introduced in Chapter 2. The question at stake in this section is whether hierarchical thematicity, furthermore, reflects better the acoustic reality than a flat theme–rheme approach.

The sentence (1) *Ever since, the remaining members have been desperate for the United States to rejoin this dreadful group*[3] is used to exemplify the thematicity–prosody correspondence using a speech sample (by spk5f) from our corpus. (1a) illustrates a flat binary theme–rheme division of this sentence.

[3] Sentence 52 in our corpus.

(1a)

Q: *What happened ever since?*
A: [*Ever since,*]T [*the remaining members have been desperate for the United States to rejoin this dreadful group.*]R

Let us consider now (1b) for illustration of hierarchical thematicity (annotated following the guidelines established in Bohnet et al. (2013)). In (1b), a total of five partitions is identified, including three spans at level 1, a specifier (SP1), theme (T1) and rheme (R1), and two embedded spans at level 2 in the rheme: a theme (T1(R1)) and a rheme (R1(R1)).

(1b)

[*Ever since,*]SP1 [*the remaining members*]T1 [*have been desperate* [*for the United States*]T1(R1) [*to rejoin this dreadful group.*]R1(R1)]R1

Figure 6.3 shows both annotation schemes for example (1) of a speech sample from our corpus (by spk5f). Normalized acoustic parameters (z-scores for F0 ('z_F0'), intensity ('z_int') and speech rate ('z_sr')) were computed from this speech sample (by spk5f) at all intervals and tiers. These z-scores are used in Figure 6.4 to represent prosody in the different thematicity partitions.

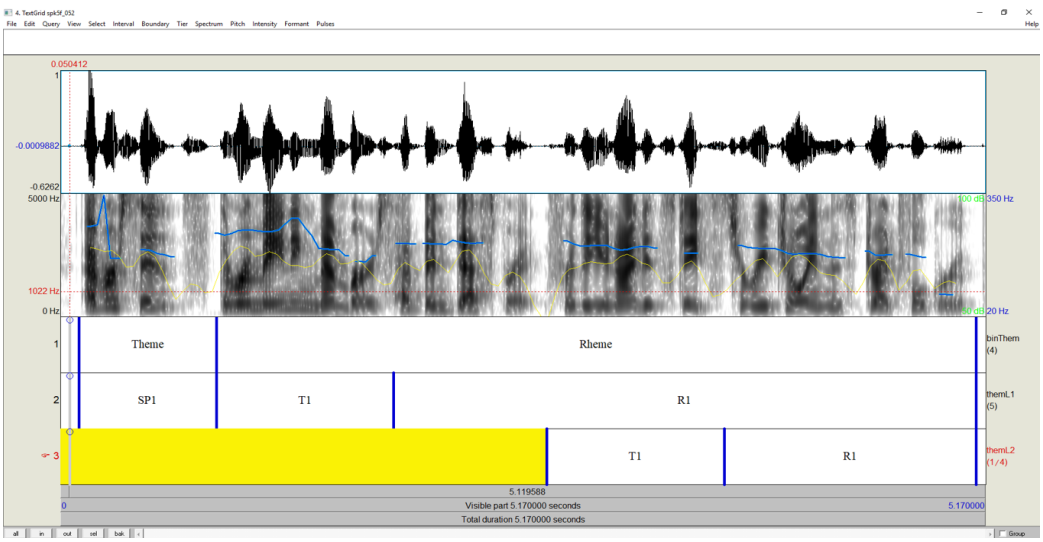
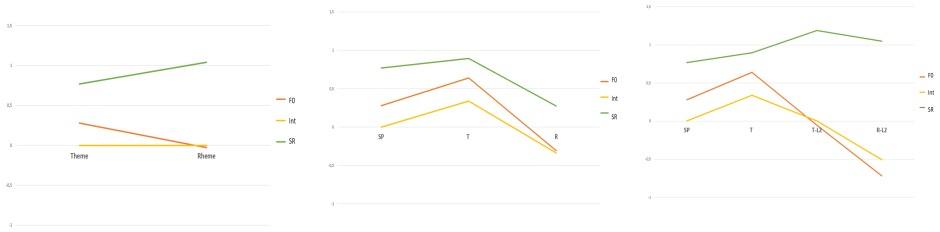


Figure 6.3: Segmentation in binary and hierarchical thematicity of (1) by spk5f.



(a) binary thematicity (b) tripartite thematicity (c) tripartite hierarchical thematicity

Figure 6.4: Acoustic parameter distribution in different thematicity partitions of (1) by spk5f.

The binary flat thematicity annotation results in the segmentation of (1) in two partitions that map onto a simple acoustic representation as depicted in Figure 6.4(a) for the example (1), as pronounced by spk5f. A tripartite division of thematicity already includes more variability in terms of prosodic parameters as shown in Figure 6.4(b). Moreover, as R1 is further subdivided into T1 and R1 at L2, these L2 divisions not only add a richer prosodic characterization of (1), as shown in Figure 6.4(c), but also suggest a clear distinction between theme and the rest of spans in all three acoustic parameters. Such a distinction is not observed in the binary representation (Figure 6.4(a)), in particular in z_int values that do not vary between theme and rheme as segmented in (1a).

Thus, a tripartite hierarchical segmentation of thematicity in this example contains more communicative spans (instead of only two in the binary representation) that better represent the acoustic variability of the speech sample by spk5f. This variability is convenient for the generation of prosodic enrichment, especially when the focus is on generating natural and expressive speech prosody.

In order to further test this correspondence, a classification experiment is designed including two datasets that contain both thematicity annotation schemes: the TRD (for the binary flat thematicity) and the HTD (for the tripartite hierarchical thematicity). These datasets include acoustic parameters extracted from one male and one female participant (spk1f and spk1m). Instances are thematicity segmentations. As there are less segmentations in the binary approach, the number of instances varies from 420 in TRD to 575 in HTD. Classes to be predicted are also different: in TRD, there are only two classes (theme and rheme), whereas in HTD, there are fifteen classes (including five thematicity partitions, i.e., T1, R1, SP1, R1-1 and R1-2, at three levels of embeddedness).

Hence, the different number of instances and classes to be predicted involves different distributions and probabilities for this classification problem. In order to compare classification results, a majority voting strategy (using a ZeroR classifier) is used as a baseline. Then, the difference from a bagging classifier (Bag) to the baseline (BL) is computed. This absolute improvement (AbsImp) over the baseline in precision, recall and f-measure is used as a metric to assess the results of the different classification tasks.

Table 6.5: Absolute improvement classification results in binary flat and tripartite hierarchical thematicity.

	Precision			Recall			F-Measure		
	BL	Bag	**AbsImp**	BL	Bag	**AbsImp**	BL	Bag	**AbsImp**
TRD	0.45	0.79	0.34	0.67	0.79	0.12	0.54	0.79	0.25
HTD	0.16	0.63	**0.47**	0.40	0.66	**0.26**	0.22	0.65	**0.43**

Results reported in Table 6.5 show that there is a greater improvement over the baseline on the HTD despite the higher number of classes included in the HTD. This result supports the argument that a hierarchical thematicity is a more appropriate description with respect to acoustic parameters of information structure than a flat theme-rheme representation, especially in long sentences with a certain amount of syntactic complexity. Despite the higher number of classes to be predicted, the distance to the baseline is greater in HTD. This implies that acoustic parameters are more homogeneously distributed over tripartite hierarchical than binary flat thematicity partitions. Thus, a tripartite hierarchical thematicity relates better to prosody in terms of acoustic parameters. In the following subsection, I explore whether including features that account for hierarchical thematicity has an effect on the prediction of ToBI labels.

The confusion matrix of the classification using the HTD (see Table 6.6) shows that a total of 86 out of 139 themes at level 1 (T1_L1) are correctly classified and the highest number of confusion occurs with specifiers at level 1 (SP1_L1) and themes at level 2 (T1_L2). In contrast, recall that in the TRD the confusion is only possible with rheme spans as there are only two labels. This demonstrates that using hierarchical thematicity, theme spans are rarely confused with rheme spans at level 1 (only three cases are confused with R1_L1). Such finding is in line with existing theories on the characteristic tunes for theme spans.

Table 6.6: Confusion matrix: prediction of thematicity in HTD.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	<– classified as
86	25	15	0	0	0	0	13	0	0	0	0	0	0	0	a = T1_L1
30	**37**	13	0	1	0	0	2	2	2	1	0	0	0	0	b = SP1_L1
3	5	**157**	1	0	0	0	0	0	0	0	0	0	0	0	c = R1_L1
3	6	8	**0**	0	0	0	2	0	0	0	0	0	0	0	d = SP2_L1
0	0	4	0	**0**	0	0	0	0	0	0	0	0	0	0	e = SP3_L1
2	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	f = R1-1_L1
0	0	2	0	0	0	**0**	0	0	0	0	0	0	0	0	g = R1-2_L1
31	4	2	0	0	0	0	**11**	1	1	0	0	0	0	0	h = T1_L2
9	5	0	0	0	0	0	2	**0**	0	0	0	0	0	0	i = SP1_L2
7	0	36	1	0	0	0	0	0	**3**	0	0	0	0	0	j = R1_L2
5	1	0	0	0	0	0	1	0	0	**0**	0	0	0	0	k = R1-1_L2
0	1	6	0	0	0	0	0	0	0	0	**0**	0	0	0	l = R1-2_L2
6	3	1	0	0	0	0	1	0	0	0	0	**0**	0	0	m = T1_L3
3	1	0	0	0	0	0	0	0	0	0	0	0	**0**	0	n = SP1_L3
0	0	10	0	0	0	0	0	0	3	0	0	0	0	**0**	o = R1_L3

6.2.2. Thematicity in the Prediction of ToBI Labels

The goal of this experiment is to assess the prediction capabilities of hierarchical thematicity, that is, whether adding attributes related to hierarchical thematicity structure improves the prediction of ToBI labels at the word level. The full dataset L2TD is used in this classification exercise. Table 6.7 shows what attributes are included in this experiment and how many distinct values each attribute contains in both the baseline (BL) and the proposed hierarchical thematicity model (HTM).

The baseline used for comparison draws upon traditional textual features, namely, PoS, syntactic dependencies and word order to predict ToBI labels, while the proposed model includes attributes that account for the description of hierarchical thematicity, namely: tripartite thematicity labels, division in propositions and embeddedness. In order to account for the overall thematicity structure within the sentence, attributes specifying the span position and the total number of spans in the sentence are also included. Regarding the fundamental aspects of prosody previously described in Chapter 2, "prosodic marker" and "prosodic type" account for the distinction of words that are prominent at the PPh or not and whether they are a pitch accent or a boundary tone.

This classification exercise contains the same number of classes (9 ToBI labels) and instances (18,792 words) for both baseline and hierarchical thematicity classification exercises. The L2T dataset is used in both experiments, but the attributes included in each prediction vary from five in the baseline to fifteen in the proposed model. The goal is to predict the ToBI label for each word and observe whether the proposed methodology that implies a hierarchical thematicity improves over the baseline. A J48 tree classifier with 10-fold cross-validation is

Table 6.7: Attributes and number of their distinct values in L2TD.

Type	Attribute	n. Distinct Values Baseline	HTM
General	Word Position	28	
	Total n. Words	22	
Syntax	Function	28	
	Dependency	36	
	PoS	31	
Thematicity	Proposition	—	7
	Embeddedness		2
	L1		6
	L2		3
	L3		2
	Total n. Spans		12
	Span Position		10
	Sentence n.		109
Prosody	Prosody Marker	—	2
	Prosody Type		4

used in both classification experiments.

Results shown in Table 6.8 yield an improvement of the proposed information structure–prosody interface over the baseline in all classes, except for precision and f-measure of HH%. A plausible explanation for this is that the label HH% represents, in this case, an end of a question intonation, and there are few questions in the corpus, and punctuation is not being taken into account for the classification exercise. However, this could be easily overcome including a rule for question mark rising intonation, as it is actually done in state-of-the-art prosody generation modules. F-measure of L* and H* increases with the proposed model by nearly 0.08 points and 0.12 points respectively. The prediction of L*+H and LH% labels, which are traditionally associated with theme spans, also improves in terms of the F-measure by 0.06 in both cases. This means that the proposed methodology is able to generate correctly more prosodic variation by means of rising bitonals and pitch accents than the baseline.

The analysis of the confusion matrices for the information structure–prosody interface (Table 6.9) and the baseline (Table 6.10) provides evidence that a prediction methodology that accounts for prosodic markers and prosodic types guarantees that errors are not made within categories that do not apply. As can be seen in the confusion matrix, L*+H is confused with H* or L* labels, that is, with other PAs labels, but never with BT, whereas the confusion matrix of the baseline shows errors within all typologies. In the baseline (see Figure 6.10), L*+H is confused,

Table 6.8: Prediction of ToBI labels using hierarchical thematicity: classification results.

	Precision		Recall		F-Measure	
	BL	HTM	BL	HTM	BL	HTM
S	0.82	**0.99**	0.89	**1**	0.85	**0.99**
L*+H	0.65	**0.69**	0.66	**0.73**	0.65	**0.94**
LL%	0.80	**0.81**	0.83	**0.87**	0.82	**0.84**
LH%	0.44	**0.50**	0.41	**0.48**	0.43	**0.49**
U	0.97	**1**	0.99	**1**	0.98	**1**
H*	0.57	**0.63**	0.48	**0.64**	0.52	**0.64**
L*	0.53	**0.59**	0.41	**0.50**	0.46	**0.54**
HL%	0.53	**0.57**	0.46	**0.49**	0.49	**0.52**
HH%	**0.90**	0.84	0.87	**0.90**	**0.88**	0.87
Average	0.78	**0.86**	0.79	**0.86**	0.79	**0.86**

apart from H* and L*, with S and U labels and even with the boundary tone LH% .

Table 6.9: Confusion matrix: prediction of ToBI labels in HTM.

a	b	c	d	e	f	g	h	i	<– classified as
5,124	0	0	0	0	0	2	0	0	a = S
0	**1,426**	0	0	0	393	123	0	0	b = L*+H
0	0	**1,458**	105	0	0	0	111	1	c = LL%
0	0	111	**276**	0	0	0	185	2	d = LH%
1	0	0	0	**5,789**	0	0	0	0	e = U
2	474	0	0	0	**1,210**	214	0	0	f = H*
1	158	0	0	0	318	**483**	0	0	g = L*
0	0	239	167	0	0	0	**387**	2	h = HL%
0	0	1	2	0	0	0	0	**27**	i = HH%

As S is the category with more instances, the baseline produces errors of S assignment in all labels that are to be predicted. In an implementation setting, there will be a prediction of a lexical accent in most of the cases, which would result in a monotonous prosody. Therefore, the improvement of the proposed model also guarantees that more variability and, consequently, more expressiveness is introduced in the resulting synthesized speech.

Taking into account that this classification exercise is done such that equal conditions are met in both the baseline and HTM, results are promising. All in all, this experiment shows significant advances of the proposed information structure–prosody interface compared to the state of the art, especially when predicting boundary tones, which are instrumental for the generation of communica-

Table 6.10: Confusion matrix: prediction of ToBI labels in BL.

a	b	c	d	e	f	g	h	i	<– classified as
4,549	193	14	20	17	223	86	24	0	a = S
270	**1,277**	1	4	50	246	88	6	0	b = L*+H
32	1	**1,396**	100	32	2	3	109	0	c = LL%
39	1	113	**237**	4	3	8	168	1	d = LH%
2	9	3	0	**5,766**	9	1	0	0	e = U
381	376	3	7	48	**914**	163	8	0	f = H*
210	115	3	11	12	208	**395**	6	0	g = L*
42	6	217	154	1	8	2	**363**	2	h = HL%
0	0	1	2	0	0	0	1	**26**	i = HH%

tive pauses and a varied range in prominent intonation in long complex sentences containing few punctuation marks.

6.2.3. From Acoustic Parameters to Thematicity

After having shown the appropriateness of the proposed methodology for generation of ToBI labels using words as the minimal unit, I set out to demonstrate the hypothesis that acoustic parameters are related to thematicity labels at the level of two different partitions: the sentence as a whole and each thematicity span. The objective of these experiments is to observe in isolation acoustic parameters and hierarchical thematicity in order to get a closer insight on their relationship. In these experiments, the correspondence of prosody and thematicity is put to test assuming a bidirectional relation between them, but acknowledging that both of them are dependent upon other linguistic phenomena.

These experiments use acoustic parameters instead of ToBI labels, because it is foreseen that results from this exercise are applicable to an analysis pipeline within a CTS system to predict thematicity structure from speech. In such a scenario, the need for automatic extraction of prosodic cues is a requirement, and consequently, the mapping from the speech signal to a ToBI representation would only introduce an unnecessary and costly intermediate step.

Prediction of Thematicity Labels

The TSD with all thematicity labels is used to perform the prediction of thematicity labels (a total of thirty-one distinct labels) using as attributes acoustic features and number of words in each span. The purpose of the experiment is to observe the correspondence between hierarchical thematicity and acoustic param-

eters using all speech samples in our corpus (by a total of twelve speakers). A ZeroR classifier is used as baseline to evaluate and compare the level of improvement considering the unbalanced nature of our corpus. Table 6.11 shows precision (P), recall (R) and f-measure (F) results for each class (using a bagging classifier) and average results for the bagging classifier and baseline (BL).

Table 6.11: Average prediction results for each class in TSD.

	Precision		Recall		F-Measure	
	BL	TSD	BL	TSD	BL	TSD
R1	0.22	**1**	**1**	0.08	**0.36**	0.15
R1(SP2)	0	**1**	0	**0.50**	0	**0.67**
R1(P4)	0	**1**	0	**0.25**	0	**0.40**
T1(P5)	0	**1**	0	**0.25**	0	**0.40**
T1(P3)	0	**0.90**	0	**0.75**	0	**0.82**
R1(T1)	0	**0.89**	0	**0.67**	0	**0.76**
T1(P4)	0	**0.86**	0	**0.50**	0	**0.63**
R1-2	0	**0.85**	0	**0.46**	0	**0.60**
T1(SP1)	0	**0.84**	0	**0.81**	0	**0.82**
R1(SP1)	0	**0.80**	0	**0.77**	0	**0.78**
R1(P2)	0	**0.78**	0	**0.69**	0	**0.73**
T1(T1)	0	**0.78**	0	**0.58**	0	**0.67**
R1(P5)	0	**0.75**	0	**0.50**	0	**0.60**
T1(R1)	0	**0.74**	0	**0.28**	0	**0.40**
R1(R1)	0	**0.72**	0	**0.29**	0	**0.42**
T1	0	**0.69**	0	**0.86**	0	**0.77**
T1(P2)	0	**0.66**	0	**0.58**	0	**0.62**
R2	0	**0.66**	0	**0.71**	0	**0.69**
SP2	0	**0.64**	0	**0.35**	0	**0.45**
SP1	0	**0.63**	0	**0.43**	0	**0.51**
R1-1	0	**0.6**	0	**0.13**	0	**0.21**
R1-1(P2)	0	**0.57**	0	**0.33**	0	**0.42**
SP1(SP1)	0	**0.50**	0	**0.25**	0	**0.33**
T1(SP2)	0	**0.25**	0	**0.17**	0	**0.20**
Average	0.05	**0.71**	0.22	**0.71**	0.08	**0.70**

The confusion matrix of the bagging classifier is shown in Table 6.12. Classes with a higher presence in the dataset are often confused when they tend to be located in the same position within the sentence, for instance, T1 is confused with SP1, that are both usually located at the beginning of the sentence. More interesting is the fact that embedded themes (T1(SP1), T1(R1), T1(P2), T1(P3) and T1(P4)) are confused with level 1 themes (T1). This indicates that themes share some acoustic properties regardless their level of embeddedness. The same occurs in embedded rheme spans (R1(SP1), R1(R1), R1(P2), R1(P4) and R1(P5)).

Results from this experiment on TSD proves a significant prediction potential of acoustic features for thematicity labels. If we compare the improvement over

Table 6.12: Confusion matrix: prediction of thematicity in TSD.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	<– classified as
1084	55	42	0	1	0	0	6	1	16	1	5	0	1	0	1	0	0	0	0	0	1	0	0	0	a = T1
37	**957**	17	6	0	2	0	16	3	2	15	1	5	1	2	0	0	0	0	0	0	0	0	0	1	b = R1
197	31	**208**	5	1	0	0	4	1	18	5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	c = SP1
1	15	23	**21**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	d = SP2
21	0	0	0	**3**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = R1-1
0	10	1	0	0	**11**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	f = R1-2
0	7	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = R2
2	24	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	h = P3
15	2	1	0	0	0	0	0	**155**	12	4	0	0	0	0	0	0	0	0	3	0	0	0	0	0	i = T1(SP1)
64	2	12	0	0	0	0	0	4	**126**	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	j = T1(P2)
11	32	3	1	0	0	0	1	3	3	**149**	0	2	0	0	0	0	0	1	0	0	0	0	0	0	k = R1(P2)
35	7	3	0	0	0	0	0	6	1	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0	l = T1(R1)
2	37	0	0	0	0	0	2	0	2	5	0	**21**	0	0	0	0	0	0	0	0	0	0	0	0	m = R1(R1)
6	3	0	0	0	0	0	0	0	1	0	0	0	**14**	0	0	0	0	0	0	0	0	0	0	0	n = T1(T1)
4	2	0	0	0	0	0	1	0	0	0	0	0	0	**16**	0	0	0	0	0	0	0	0	0	0	o = R1(T1)
2	0	0	0	0	0	0	0	1	2	0	0	0	0	0	**4**	1	0	0	2	0	0	0	0	0	p = R1-1(P2)
3	0	4	0	0	0	0	0	0	1	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	q = SP1(SP1)
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	r = SP1(P2)
1	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	**18**	0	0	0	0	0	0	s = T1(P3)
2	0	0	0	0	0	0	0	7	0	0	0	0	0	0	1	0	0	0	**2**	0	0	0	0	0	t = T1(SP2)
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	0	u = R1(SP2)
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	0	0	0	v = T1(P4)
0	8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**3**	0	0	w = R1(P4)
7	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**3**	0	x = T1(P5)
1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**6**	y = R1(P5)

the baseline classifier, a considerable increase in all measures is attained (P=+0.64, R=+0.49, F=+0.62). Moreover, a precision of 1 or higher than 0.85 is achieved for eight labels (R1, R1(SP2), R1(P4), T1(P5), T1(P3), R1(T1), T1(P4) and R1-2), half of which involve theme spans and embeddedness. This further supports the argument that on the one hand, themes have distinct prosodic characteristics, as previously suggested in the literature, and, on the other hand, that hierarchical thematicity is a more appropriate representation of information structure than traditional approaches.

Prediction of Thematicity at Sentence Level

A final experiment is carried out at the sentence level. For each sentence span, acoustic parameters are extracted to predict the thematicity label sequence at L1 using the SSD (a total of seventeen distinct labels are to be predicted). A simple rule classifier (ZeroR), based on a majority vote, is used as baseline. Classification with ZeroR shows a low average scores on precision (P=0.29), recall (R=0.54) and f-measure (F=0.38). Then, a bagging classifier is used and results show a considerable increase in all measures with an average absolute improvement over the

baseline of P=+0.44 R=+0.19 and F=+0.36. Table 6.13 reports precision, recall and f-measure results from this classification.

Table 6.13: Average prediction results (P, R and F) for each class in SSD.

	Precision		Recall		F-Measure	
	BL	SSD	BL	SSD	BL	SSD
T1R1(P2)T1R1(P3)	0	**1**	0	**1**	0	**1**
R1	0	**0.82**	0	**0.86**	0	**0.84**
T1R1	0.54	**0.80**	**1**	0.88	0.70	**0.83**
R1SP1	0	**0.77**	0	**0.72**	0	**0.74**
R1T1	0	**0.75**	0	**0.75**	0	**0.75**
SP1SP2R1	0	**0.74**	0	**0.58**	0	**0.65**
SP1SP2T1R1	0	**0.73**	0	**0.46**	0	**0.56**
T1(P2)R1(P3)	0	**0.67**	0	**0.83**	0	**0.74**
SP1T1R1	0	**0.67**	0	**0.61**	0	**0.64**
T1R1SP1(P2)T1R1(P3)	0	**0.62**	0	**0.54**	0	**0.58**
T1R1SP1	0	**0.60**	0	**0.49**	0	**0.54**
SP1T1SP2R1	0	**0.51**	0	**0.33**	0	**0.42**
R1-1T1R1-2	0	**0.50**	0	**0.42**	0	**0.46**
T1SP1R1	0	**0.46**	0	**0.50**	0	**0.48**
SP1T1R1R2	0	**0.43**	0	**0.25**	0	**0.32**
SP1R1-1T1R1-2	0	**0.17**	0	**0.08**	0	**0.11**
Average	0.29	**0.73**	0.54	**0.75**	0.38	**0.74**

The confusion matrix analysis shows that the most common errors fall within the T1R1 class, especially with the initial (SP1T1R1) and final (T1R1SP1) specifier thematicity label sequence. If we consider that thematicity segmentation is highly dependent on syntactic dependencies, these results lead us to think that adding linguistic features to the classification problem, precision, recall and f-measure would reach very high scores. However, such an experiment is out of the scope of the present work as we want to test the relation of prosody to information structure independently form any other layer, as stated before. In any case, good precision results are attained for sentences containing two coordinated propositions (T1R1(P2)T1R1(P3), P=1) and rhematic sentences (R1, P=0.82).

6.2.4. Discussion

To sum up, this section provides evidence from our corpus that: (i) hierarchical thematicity is a better representation than a simple theme–rheme division; (ii) the prediction of ToBI labels improves using thematicity; (iii) and it is possible to predict the thematicity label using only acoustic features as attributes. All these

Table 6.14: Confusion matrix: prediction of thematicity in SSD.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<– classified as
618	27	4	19	6	4	6	0	1	4	2	4	2	4	5	0	a = T1R1
73	**115**	0	10	1	0	8	0	0	0	0	0	0	0	0	5	b = SP1T1R1
7	0	**4**	1	0	0	0	0	0	0	0	0	0	0	0	0	c = SP1T1SP2R1
26	12	0	**44**	0	0	0	0	0	0	0	2	0	0	0	0	d = T1R1SP1
5	0	0	0	**26**	0	0	0	5	0	0	0	0	0	0	0	e = R1T1
8	1	0	0	0	**1**	2	0	0	0	0	0	0	0	0	0	f = SP1R1-1T1R1-2
5	11	0	0	0	1	**28**	0	0	0	0	0	0	0	0	3	g = T1R1SP1(P2)T1R1(P3)
0	0	0	0	0	0	0	**12**	0	0	0	0	0	0	0	0	h = T1R1(P2)T1R1(P3)
2	0	0	0	3	0	0	0	**30**	0	0	1	0	0	0	0	i = R1
9	1	0	0	0	0	1	0	0	**1**	0	0	0	0	0	0	j = SP1T1R1R2
5	0	0	0	0	0	0	0	0	0	**5**	0	2	0	0	0	k = R1-1T1R1-2
9	0	0	3	0	0	0	0	2	0	0	**21**	0	1	0	0	l = R1SP1
10	0	0	0	0	0	0	0	0	0	3	0	**11**	0	0	0	m = SP1SP2T1R1
7	0	0	0	0	0	0	0	0	0	0	3	0	**14**	0	0	n = SP1SP2R1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	o = T1(P2)R1(P3)
0	3	0	0	0	0	3	0	0	0	0	0	0	0	0	**6**	p = T1SP1R1

evidences support the argument that there is a strong correlation between hierarchical thematicity and prosody.

A hierarchical thematicity structure has been shown to correlate better with intonation labels (using the ToBI convention as previous studies do for the representation of prosody) than binary flat thematicity. However, such a correlation still does not solve the problem of a one-to-one mapping between a specific intonation label (e.g., the ToBI label H*) and a static acoustic parameter (e.g., an increase of 50% in fundamental frequency). Besides, the ToBI labeling involves a considerable amount of manual annotation efforts by trained experts.

In this section, the proposed methodology for the information structure–prosody interface has been examined using an empirical approach by means of statistical tests and classification experiments. However, it still remains to be demonstrated how this empirical data can be applied to actually generate prosody for expressive speech synthesis. Next section addresses this issue.

6.3. How do we Get to Expressive Prosody Generation?

In the previous section, classification experiments were aimed at proving the appropriateness of a hierarchical thematicity approach. However, it was still not clear how these empirical findings match the technical requirements for prosody

generation. In the present section, I embark on the exploration of how to get to prosody generation within the framework of the information structure–prosody interface.

In what follows, I set out to assess whether ToBI labels can be decomposed into three acoustic elements (intensity, pitch and duration) using empirical evidence from our corpus. To do so, the full dataset ALD is used. The prediction of ToBI labels carried out using only linguistic features in Section 6.2 is further expanded to acoustic elements, namely intensity, F0 and duration. If prediction results when combining linguistic and acoustic features are better than using only linguistic features, this would imply that the proposed representation of ToBI is suitable for mapping to a variety of acoustic elements, and not only to F0 contours.

For the experiments in this section, a J48 tree classifier with a 10-fold cross-validation is used. In order to account for the linear nature of the classification problem, a word identification number is used to mark the position of the word within each sentence. Accuracy (A), kappa (k) and root mean square error (RMSE) are reported to assess the performance of a combination of linguistic and acoustic features. The full dataset ALD is used in these experiments to observe the different prediction results involving a combination of attributes. ALD has a total of 20 attributes distributed in linguistic and acoustic levels and different elements as Table 6.15 shows.

As already mentioned, ALD contains 18,792 instances that correspond to word segments of speech samples from all twelve participants and 109 sentences. The prediction involves nine classes of ToBI labels at word segments with the average distribution reported in Table 6.16.

6.3.1. Testing Acoustic Parameters

The fist experiment on the ALD is done selecting the acoustic level and testing the prediction capabilities of the different acoustic elements, namely, intensity, F0 and duration. Table 6.17 shows that when F0, intensity and duration are used on their own to predict prosodic labels, the best accuracy score is achieved by duration (A=54%) and the lowest by F0 and intensity (A=44%). The combination of all three elements (i.e., intensity, duration and F0) achieves the same accuracy as using only duration cues (A=54%). This suggests that our approach for ToBI annotation is accounting not only for F0 movements, but also for a combination of three prosodic elements in our corpus of read speech.

Table 6.15: Combination of attributes in ALD.

Level	Element	Attribute
Linguistic	Position	Word Position
		Total n. Words
	Syntax	Function
		Dependency
		PoS
	Thematicity	Proposition
		Embeddedness
		L1 Thematicity
		L2 Thematicity
		L3 Thematicity
		Total n. Spans
		Span Position
Acoustic	Intensity	Intensity Range
		Minimum Intensity
		Mean Intensity
		Maximum Intensity
	F0	F0 Range
		Minimum F0
		Maximum F0
	Duration	Word Duration

Taken together, these results indicate that, as already pointed out by Audibert et al. (2005), in order to obtain a more natural synthesized voice, more acoustic elements (apart from F0) should be integrated.

6.3.2. Adding Linguistic Features

Following the experiment in Section 6.2 that used linguistic features (including thematicity) to predict ToBI labels, I set out to explore now to what extent the prediction of ToBI labels improves when acoustic and linguistic features are combined in the full dataset ALD. Table 6.18 presents the results from combining the acoustic level with each linguistic element: word position, syntax and thematicity; and the linguistic level with each acoustic element: intensity, F0 and duration.

Compared to the previous experiment, where only the acoustic level (where A=56% was achieved in the best scenario) was used, prediction accuracy improved considerably when the acoustic level is combined with linguistic elements. Syntactic attributes are particularly useful (leading to an accuracy of 72%). The picture improves even further when the linguistic level is kept static and individual acoustic elements (or a combination thereof) are used for classification. In particular, the tandem linguistic element and F0 reaches A=78%. All in all, the combination of acoustic and linguistic features improves precision by 23 points

Table 6.16: Distribution of classes in ALD.

Class	n. of Instances	Distribution
U	5790	31%
S	5126	27%
L*+H	1942	10%
H*	1900	10%
LL%	1675	9%
L*	960	5%
HL%	795	4%
LH%	574	3%
HH%	30	0.2%

Table 6.17: Prediction of ToBI labels from acoustic parameters: classification results.

Acoustic Elements	A	k	RMSE
Intensity + Duration + F0	53.66%	0.41	0.29
Duration + Intensity	53.41%	0.41	0.29
Duration + F0	53.15%	0.40	0.29
Duration	53.69%	0.39	0.26
F0 + Intensity	43.77%	0.28	0.32
Intensity	41.44%	0.24	0.31
F0	39.47%	0.22	0.31

compared to using only acoustic elements.

Nevertheless, it needs to be underlined that the best result is achieved by the combination of the linguistic level with each acoustic element separately (A=77%), rather than by the combination of all linguistic and acoustic features (when 73% is reached). According to preliminary experiments on speaker specific datasets in Subsection 6.1.2, ToBI labels represent a different combination of acoustic elements. Moreover, the F0 element showed a statistically significant differences across speakers. Consequently, it makes sense that when the full dataset (including twelve speakers) is used in this classification exercise, F0 yields better results.

Results suggest that prosodic labels show a characteristic combination of acoustic parameters. Hence, a description of ToBI labels in terms of acoustic parameters is introduced below. Figure 6.5 shows that both prosodic labels and prosodic marks (PA, BT, S and U) are represented as a distinct combination of intensity, F0 and duration elements. Thus, PAs (H*, L*, L*+H) are characterized by positive or null deviation in all three acoustic elements, while words labeled as 'S' have

Table 6.18: Combining linguistic and acoustic elements for ToBI prediction

Attributes		Accuracy	Kappa	RMSE
Acoustic Level +	Position	0.59	0.48	0.28
	Thematicity	0.60	0.49	0.27
	Syntax	0.72	0.64	0.23
Linguistic Level +	Intensity	0.77	0.70	0.20
	Duration	0.77	0.70	0.20
	F0	**0.78**	**0.72**	**0.20**
All attributes		0.73	0.65	0.22

little or no deviation at all. BTs, on the other hand, are characterized by a high positive deviation in duration and negative deviation in intensity and F0. Finally, words labeled as 'U' have outstandingly low negative deviation in duration and negative deviation in intensity and duration.

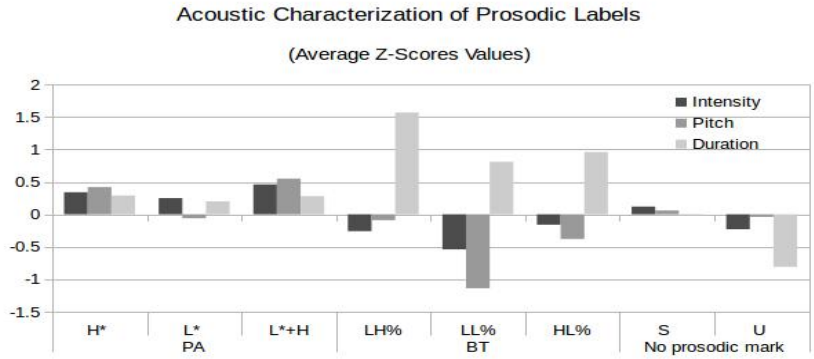


Figure 6.5: Characterization of ToBI labels combining acoustic elements.

This characterization is instrumental for prosody enrichment in TTS applications using ToBI labels as the annotated labels need to be mapped to the acoustic signal. However, it must be taken into account that previous experiments in Section 6.1 demonstrated that some speech samples vary in the way ToBI labels are mapped to acoustic parameters. Consequently, in an implementation scenario of speech synthesis, it can be foreseen that prosody labels may be mapped onto a varied range of acoustic parameters that guarantee variability and deal with monotony in synthetic voices.

6.4. A Corpus-driven Analysis of the Thematicity–Prosody Correspondence

The analysis introduced in this section aims to cover the two main goals of this dissertation: providing empirical evidence on the correlation between prosody and hierarchical thematicity and establishing a ground that can serve to derive prosodic tags for implementation in a CTS application. The characterization presented in Subsection 6.4.1 comprises the manually ToBI annotated working corpus. The advantage of this representation is that it will serve to link previous theoretical constructs on the information structure–prosody interface to the present empirical approach, but, as aforementioned, this annotation cannot be reliably derived using state-of-the-art automatic tools like AuToBI. To overcome this shortcoming, Subsection 6.4.2 presents a characterization based on the automatic extraction and computation of acoustic parameters using the modular prosody tagger presented in Chapter 5.

6.4.1. Correspondence between Thematicity and Manual ToBI Annotation

This section reports on the correspondence between manually annotated ToBI and hierarchical thematicity. This analysis provides information about the level of coincidence in prosodic phrases, compared to thematicity spans and the commonest intonation patters of T1, R1, and SP1 at L1 and L2.

I commence by reporting the level of coincidence in hierarchical partitions and prosodic phrases (PPh) taking into account partial matches, namely, partitions that coincide only on one end (i.e., the final word is the same in both PPh and thematicity span) and full matches, i.e., both beginning and ending words are the same in PPh and thematicity spans. The total amount of PPhs coinciding with the final boundary of a thematicity span in our corpus is 72%. However, if we do not count the end of a sentence as a boundary neither for thematicity nor for PPh, the rate of coincidence goes down to 51%. If we restrict coincidence rates to L1 thematicity with the tripartite division (theme–rheme–specifier) proposed by Mel'čuk, the coincidence of thematicity spans and PPhs in our corpus decreases to 56%. If we only consider coincidence in theme partitions at L1, only 23% of the PPh coincide with level 1 theme span divisions. Consequently, coincidence in prosodic phrasing and thematicity spans is positively correlated with the amount of levels.

In Section 6.2, quantitative evidence was presented that supports the argument

that a tripartite hierarchical thematicity structure captures better intonation contours than a flat theme–rheme structure. As further evidence, qualitative data is provided in the present section. Taken together, such evidence confirm results presented in Section 6.2 and allows us to distill detailed findings concerning the correlation between thematicity at different levels of embeddedness and intonation, namely:

- Main themes (i.e., T1) are typically characterized by a rising intonation; either an L*+H PA or, if they contain a full PPh, a final LH% BT. However, a falling final tone HL% or LL% is also found in our corpus. Example (3) shows the annotation of thematicity for the sentence *The investment choices offered by the pension fund currently are limited to a stock fund , an annuity and a money - market fund*[4].

  (3)

  [**The investment choices offered by the pension fund**]T1 [*currently are limited to a stock fund, an annuity and a money - market fund*]R1.

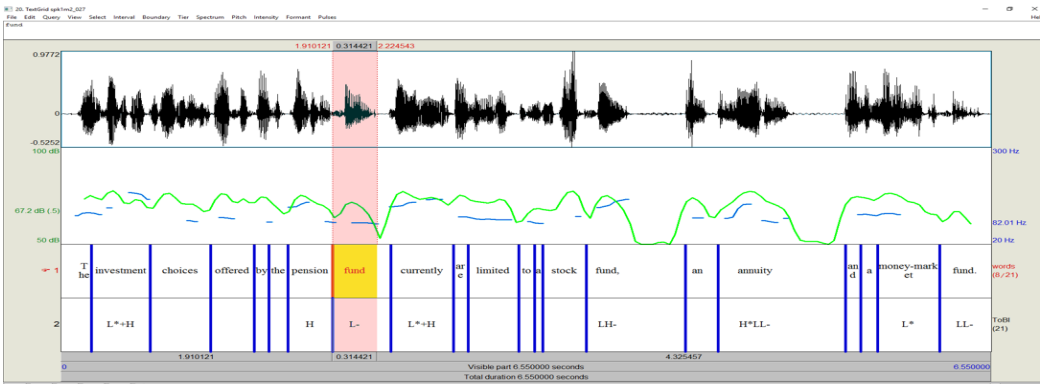


Figure 6.6: T1 as L*+H HL% intonation in (3) by spk1m.

In example (3), a rising PA on *investment* is common across speakers from different dialectal areas (included in the previous comparison in Section 6.1) and all five speakers make a pause at the end of the theme span *pension fund*. The BT in three cases is LH%, while, in the two other cases, the speakers chose a falling or flat BT, as shown in Table 6.19. It is noteworthy that the speakers' decision to make a rising BT at the end of the theme span triggers a flat of falling BT in the next PPh (see spk1f and spk2m in Table 6.19), while those who chose a falling or flat BT at the end of T1 are producing a

[4]Sentence 27 in our corpus.

Table 6.19: ToBI annotation of (3) as read by five participants.

	T1		R1				
	investment	pension fund	currently	stock fund,	annuity	money	fund
spk1f	L*+H	LH%	L*+H	**HL%**	H* LL%	L*	LL%
spk1m	L*+H	**HL%**	L*+H	LH%	H* LL%	L*	LL%
spk2m	L*+H	LH%	L*+H	**LL%**	H* LH%	L*	LL%
spk4m	L*+H	**LL%**	L*+H	**HL%**	H* LL%	L*	LL%
spk5f	L*+H	LH%	L*+H	LH%	L* LH%	L*	LL%

rising and falling intonation respectively (see spk1m and spk4m Table 6.19). All of them, however, coincide in the flat final contour (L* LL%) at the end of the sentence.

- Embedded themes (e.g., T1(T1)) often present L*+H tones, which conclude with either an HL% or LL% boundary tone depending on the context, regardless whether they are embedded in T1 or in R1. Example (4) shows the annotation of thematicity for the sentence *What triggered the latest clash was a skirmish over the timing of a New Zealand government bond issue*[5] including an embedded theme, i.e. T1(T1) highlighted in bold.

(4)

[*What* [*triggered*]R1(T1) [**the latest clash**]T1(T1)]T1 [*was a skirmish over the timing of a New Zealand government bond issue*]R1.

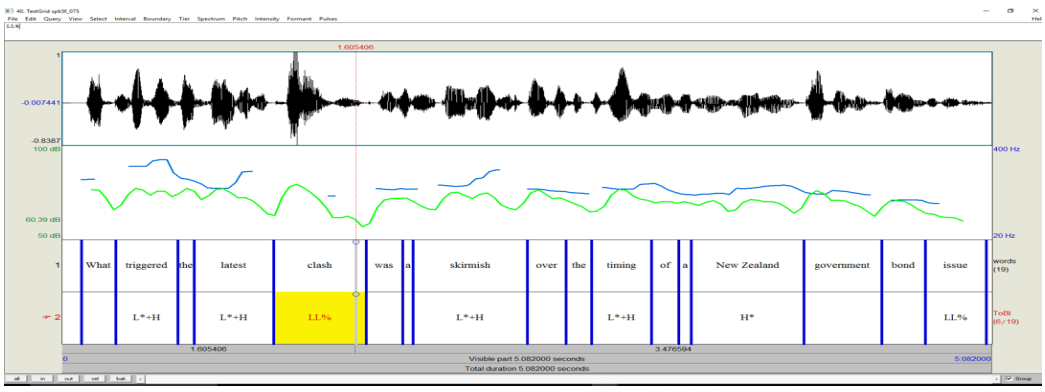


Figure 6.7: T1(T1) as L*+H LL% intonation in (4) by spk5f.

Table 6.20 shows an overall coincidence in example (4) signaling the T1(T1) with a BT. Most of the samples show a falling tune HL%. However, it is

[5]Sentence 75 in our corpus.

Table 6.20: ToBI annotation of (4) as read by five participants.

	T1			R1			
L2	R1(T1)	T1(T1)					
	triggered	latest	clash	skirmish	timing	New Zealand	bond issue.
spk1f		L*+H	HL%	L*+H			HL%
spk1m	L*+H		HL%	H*	L*+H		HL%
spk2m	H*		LH%	L*+H	L*+H		LL%
spk4m	L*+H		HL%	H*	L*+H	H*	L* LL%
spk5f	L*+H	**L*+H**	**LL%**	L*+H	L*+H	H*	LL%

true, that in this case, the end of T1(T1) coincides with the end of T1. In the example of spk5f (see Figure 6.7), T1(T1) contains a rising PA L*+H.

In this annotation, large prosodic phrases are labeled without considering smaller units. In most cases the PA of this PPh is carried by the word *triggered*. Spk5f shows a relevant greater prominence also on this word, however, due to a significant prominent PA on *latest*, a secondary L*+H is placed there.

- At any level of theme embeddedness, monosyllabic words, especially personal pronouns, are expected to present prosodic characteristics that may not be represented in the proposed ToBI annotation, such as forming bigger prosodic units with immediately preceding and subsequent words. I do not aim to provide here a detailed description of what prosodic processes are involved in monosyllabic words. Nonetheless, there are some interesting facts, especially concerning monosyllabic personal pronouns (which are 'given' in terms of Mel'čuk, and therefore do not carry prosodic marking) located at theme spans, which are worth mentioning as they affect and may even change the prosodic patterns found in these spans. Example (5) shows the annotation of thematicity for the sentence *Men who have played hard all their lives aren't about to change their habits, he says*[6]. (5) contains three types of monosyllabic themes highlighted in bold below: *Men* in T1; and at L2, *who* as theme of the embedded proposition T1(P2); and *he* as theme of the specifier T1(SP1).

(5)

[**Men** {[**who**]T1(P2) [*have played hard all their lives*]R1(P2)}P2]T1
[*are n't about to change their habits*]R1 , [[**he**]T1(SP1) [*says.*]R1(SP1)]SP1

[6]Sentence 96 in our corpus.

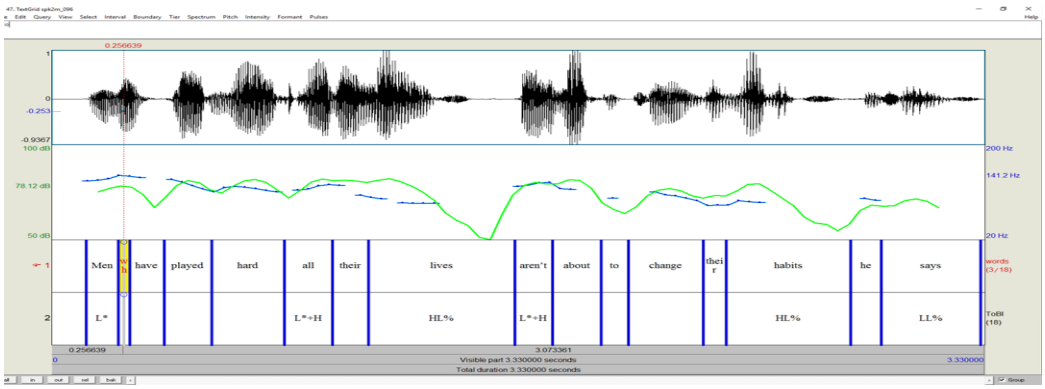


Figure 6.8: Monosyllabic themes in (5) by spk2m.

Table 6.21: ToBI annotation of (5) as read by five participants.

L2	T1	P2		R1			SP1	
	Men	all	lives	aren't	change	habits,	T1 he	R1 says.
spk1f	L*+H	L*	LH%	L*+H		HL%	LL%	
spk1m	L*+H	H*	LH%	L*+H		HL%	LL%	
spk2m	**L***	L*+H	HL%	L*+H		HL%	LL%	
spk4m	L*+H	L*+H	**HL%**	H*	L*	LL%	LL%	
spk5f	L*+H	L*+H	**HL%**	H*	L*	LL%	LL%	

The monosyllabic main theme (T1) *Men* in example (5) carries the characteristic bitonal L*+H, except for sample spk2m (see Figure 6.8): here, the speaker forms the prosodic unit *Men who have played hard*, which is pronounced at a higher speech rate than other samples. This derives in a change of prominence from the word *Men* to *all*. Regarding personal pronouns at L2 (*who* and *he*), they are unstressed in all samples.

- Specifiers at L1 (SP1) in initial positions usually involve a rising tune L*+H LH%, as shown in Table 6.22. Example (6) includes the annotation of thematicity for the sentence *On a commercial scale, the sterilization of the pollen–producing male part has only been achieved in corn and sorghum feed grains*[7].

(6)

[**On a commercial scale**]SP1 , [*the sterilization of the pollen - producing male part*]T1 [*has only been achieved in corn and sorghum*

[7]Sentence 63 in our corpus.

feed grains]R1 .

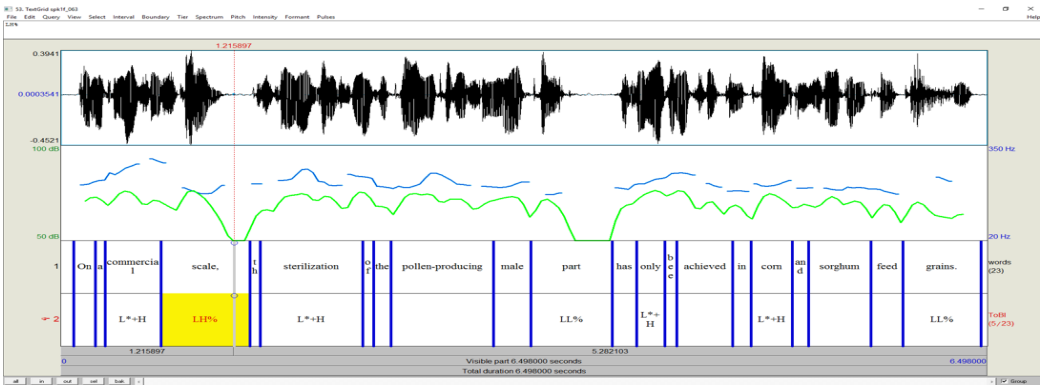


Figure 6.9: SP1 as L*+H LH% in (6) by spk1f.

Table 6.22: ToBI annotation of (6) as read by five participants.

	SP1		T1			R1			
	commercial	scale	sterilization	pollen–producing	male part	only	corn	sorghum	feed grains.
spk1f	L*+H	LH%	L*+H		LL%	L*+H	L*+H		LL%
spk1m	H*	LH%	H*		HL%	L*+H	L*+H	H*	LL%
spk2m	H*	LH%	H*		HL%	L*+H	L*+H	H*	LL%
spk4m	**L***	**LL%**	H*	**L*+H**	HL%	L*+H	L*+H		LL%
spk5f	L*+H	LH%	H*	**L*+H**	HL%	L*+H	**L*+H**		LL%

However, when SP1 contains an embedded rheme R1(SP1), a flat contour L* LL% is the characteristic ToBI pattern. In our corpus, this kind of specifiers mostly coincides with reported speech, and are located either at the end of the sentence or in the middle of other spans. Example (5) previously showed this type of flat pattern (see Table 6.21). In case a specifier contains a longer T1(SP1), and is located in initial position, specifiers involve either a falling or a rising intonation. Example (7) shows the annotation of thematicity for the sentence *As Yogi Berra might say, it's deja vu all over again*[8].

(7)

[[**As**]SP1(SP1)[**Yogi Berra**]T1(SP1) [**might say**]R1(SP1)]SP1 , [*it*]T1 [*'s deja vu all over again.*]R1

[8]Sentence 85 in our corpus.

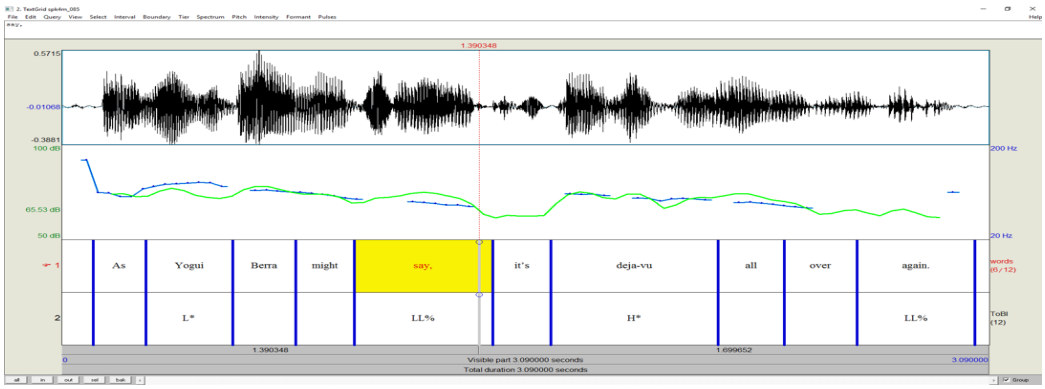


Figure 6.10: R1(SP1) as L* LL% in (7) by spk4m.

Table 6.23: ToBI annotation of (7) as read by five participants.

	SP1			T1	R1	
	Yogui	Berra	say	it	deja-vu	again.
spk1f		H*	LL%		H*	LL%
spk1m		H*	LL%		H*	LL%
spk2m	H*		LH%		H*	LL%
spk4m	**L***		**LL%**		H*	LL%
spk5f	H*		LH%		H*	LL%

Table 6.24 summarizes the most characteristic intonation patterns in main (or level 1) and embedded (or level 2) spans (L1 and L2 respectively).

Table 6.24: ToBI patterns and their associated hierarchical thematicity spans.

		T1		R1		SP1
L1		L*+H HL%	L*+H LH%	H* LL%		L*+H LL%
L2	T1	L*+H LL%	H* LL%	L*+H LL%	H* LL%	L*+H LH%
	R1	H* LL%		H* LL%		L* LL%

6.4.2. Distribution of Acoustic Parameters in Hierarchical Thematicity

This section provides a description of hierarchical thematicity in terms of relative acoustic parameters. The main goal of such a characterization is two-fold: firstly, to introduce a characterization of hierarchical thematicity that includes pa-

rameters representative of three prosodic elements and, secondly, to propose a methodology that is scalable for the generation of a thematicity-based prosodic contours using SSML tags within a CTS application. In this vein, the use of relative acoustic parameters and multiple prosodic elements is expected to contribute to a wider variability in the generation of more expressive prosodic contours in synthesized speech that will be exemplified in the next section.

Distribution of Acoustic Parameters in Thematicity Spans

Given that no significant differences have been detected between speakers in relative acoustic parameters at the sentence level, I set out to test whether thematicity spans involve different acoustic distributions. The analysis of mean normalized values of intensity (z_int), F0 (z_F0) and speech rate (z_sr) is presented in several figures grouped in terms of L1 thematicity (see Table 6.25), propositions (see Table 6.26) and, finally, theme and rheme spans across different levels of embeddedness (see Table 6.27).

Table 6.25: Distribution of acoustic parameters in L1 thematicity.

L1 Thematicity	z_int	z_F0	z_sr
T1	0.16	0.47	0.95
R1	-0.04	-0.17	0.30
R1-1	0.15	0.72	-0.20
R1-2	-0.12	**-0.39**	0.37
SP1	-0.07	0.15	1.76
SP2	-0.38	-0.55	1.05

In the analysis of L1 thematicity spans, Table 6.25 shows that there are distinct distribution patterns between T1, R1 and SP1 spans. T1 shows positive deviations in all z_int, z_F0 and z_sr parameters; R1 has negative z_int and z_F0 and positive (but lower than T1) z_sr; and SP1 is characterized by negative z_int (like R1), positive (but lower than T1) z_F0 and positive (higher than T1) z_sr. If we look at similarities in the patterns, it is noted that R1, R1-2 and SP2 show similar distribution patterns of all acoustic parameters, but they differ in the range of values.

The analysis of the propositions presented in Table 6.26 shows a clearly different characterization between all propositions, especially from P3 to P5, which present negative scores for z_int and z_F0. Moreover, P5 is substantially different in negative z_sr.

Table 6.26: Distribution of acoustic parameters in propositions and specifiers.

Propositions	z_int	z_F0	z_sr	Specifiers	z_int	z_F0	z_sr
P1	0.01	0.02	0.19	SP1	-0.07	0.15	**1.76**
P2	0.06	**-0.11**	0.83	SP1(P2)	0	0.63	**3.36**
P3	**-0.14**	**-0.29**	0.34	SP1(SP1)	0.25	-0.06	**3.13**
P4	**-0.38**	**-0.17**	1.11	SP2	-0.38	-0.55	**1.05**
P5	**-0.11**	**-0.19**	**-1.35**				

Specifiers (see Table 6.26) show differences regardless their level or span of embeddedness. Nevertheless, they all share a common feature in high positive values of z_sr.

Table 6.27: Distribution of acoustic parameters in embedded themes and rhemes.

Themes	z_int	z_F0	z_sr	Rhemes	z_int	z_F0	z_sr
T1	0.16	0.47	0.95	R1	-0.04	-0.17	0.30
T1(P2)	0.23	0.22	3.51	R1(P2)	-0.04	-0.35	0.6
T1(P3)	-0.15	-0.05	-3.56	R1(P3)	-0.67	-1.06	-0.72
T1(P4)	0.37	-0.10	0.04	R1(P4)	-0.61	-0.21	1.50
T1(P5)	-0.03	0.18	1.18	R1(P5)	-0.15	-0.47	-1.42
T1(R1)	0.12	0.21	1.30	R1(R1)	-0.23	-0.41	0.48
T1(SP1)	-0.13	-0.07	2.25	R1(SP1)	-0.21	-0.21	1.11
T1(SP2)	-0.08	-0.72	6.62	R1(SP2)	-0.83	-1.46	-1.07
T1(T1)	0.27	0.57	0.23	R1(T1)	0.27	0.80	0.89

Likewise, themes and rhemes show a diverse acoustic characterization with respect to their level and type of embeddedness with some commonalities. Table 6.27 shows distinct negative values of z_int and z_F0 with high positive z_sr in themes that are embedded in specifier spans. It is also worth mentioning that embedded themes in propositions (from P3 to P5) present negative values especially of z_sr. A more detailed study on themes based on their word length is presented in the next section.

Table 6.27 presents the acoustic characterization of rhemes. In general, rhemes share a pattern of negative intensity and F0 values and positive SR in both z-scores relative to the whole sound file and previous spans (i.e., z_int, z_F0, z_sr, z_int_p, z_F0_p, z_sr_p). The most outstanding exception is found in rhemes embedded in a theme span (i.e., R1(T1)). In fact, the acoustic characterization of R1(T1) is similar to T1 and T1(T1) (see Table 6.27) as all these spans share positive values in all parameters. However, R1(T1) differs in that it shows higher z_F0 value,

maxF0_t and z_dur_p values and slightly lower z_F0_p that T1 and T1(T1) spans.

Summing up, statistical analysis tests show significant differences between labels with respect to acoustic parameters extracted at each thematicity span. A qualitative analysis of this multidimensional parametric representation of prosody in thematicity spans proves consistent similarities and differences in labels depending on their level of embeddedness and type of span. Differences in the distribution of values are observed in all thematicity elements. This indicates that there is a wide range in the variation and characterization of thematicity spans that can be exploited in generation of more expressive synthesized speech. In Section 6.5, this characterization is used in a data-driven approach for a thematicity-based generation of SSML prosody control tags.

As theme spans in the corpus vary considerably in the number of words (from one to sixteen), next section provides an analysis of themes according to their length in words.

Distribution in Theme Spans

In what follows, I present a more detailed analysis on theme spans based on their word length to test whether word length affects the acoustic characterization of theme spans. In our corpus, more than half of the themes (55%) consist of one or two words, 38% have between three and eight words, and 8% have more than nine words. Thus, there is a diversity of themes depending on their word length. This suggests that complex syntactic structures and length in the number of words affects the overall prosodic characteristics of spans, in particular themes, which are, as aforementioned, proven to relate to distinct intonation contours.

Table 6.28: Distribution of acoustic parameters in themes with respect to their number of words.

n. of Words	z_int	z_F0	z_sr	n. of Words	z_int	z_F0	z_sr
1	-0.06	0.20	4.15	8	0.16	0.23	0.29
2	0.25	0.64	0.11	9	-0.04	0.10	0.08
3	0.25	0.38	0.18	10	0.12	0.38	0.77
4	0.28	0.35	-0.27	11	0.15	0.18	-1
5	0.26	0.48	0.61	12	-0.02	0.21	-0.50
6	0.21	0.27	0.25	14	0.09	0.06	-0.53
7	0.10	0.12	0.93	16	-0.15	0.02	-0.63

Theme spans were selected regardless their level of embeddedness and their

number of words was used as grouping factor to conduct a one-way ANOVA test. Results show there are statistically significant differences with respect to the number of words. Descriptive tests on the acoustic characterization presented in Table 6.28 show that the main differences in acoustic characterization of themes depending on their word length are found in one word themes. These single word spans are usually personal subject pronouns, and as expected in monosyllabic words for phonological reasons (personal pronouns are often lexically unstressed), they show negative z_int values and very high positive z_sr.

On the other hand, themes that have more than eleven words present a significantly low value of z_sr. However, it should be noted that themes with four words or more show duration scores higher than 1.5 relative to the previous span (i.e., z_dur_p), with low values of z_sr, even negative z_sr in the case of four- and six-word spans. These results show that the speech rate of theme spans is slower than the average values for speech rate in other spans. Consequently, long theme spans are pronounced at a slower pace respective to the speaker's average speech rate in our corpus of read speech. It is beyond of the scope of this dissertation to analyze syllables length, but in order to draw definite conclusions on thematic monosyllabic pronouns, empirical data should be collected for pronouns in different thematicity spans to be able to compare them properly.

6.5. Thematicity-based Speech Synthesis Experiments

This section addresses the main objective of this dissertation: bridging the gap between theoretical studies on the information structure–prosody interface and its implementation in CTS applications. Subsection 6.5.1 points out the main shortcomings of TTS applications due to the lack of communicative structure for prosody prediction. To this aim, I briefly exemplify common failures in prosody generation by comparing two TTS systems with different voice qualities. Then, I present the proposed thematicity-based prosody enrichment in Subsection 6.5.2 as a module for a CTS application that automatically generates hierarchical thematicity within the NLTG pipeline. The evaluation of the system is carried out by means of perception tests and objective metrics in Subsection 6.5.3. Finally, conclusions are drawn in Subsection 6.5.4.

6.5.1. The Lack of Communicative Structure in TTS Applications

In this section, several problems that have not been considered so far in the application of the information structure–prosody interface are presented. The first problem goes beyond voice quality in TTS applications, and highlights the fact that communicative structure is neglected for prosody generation in TTS applications independently of which speech generation technique is used. Two TTS systems are presented: MaryTTS[9] and Bluemix[10]. The analysis of different speech synthesis techniques is introduced, then a comparison to gold standard is made to underline the differences in communicative prosody between human and synthesized speech samples.

The second issue, presented in this section is more of a methodological problem in applying ToBI labels for the generation of thematicity-based prosody contours. I explore the possibility of using ToBI labels in MaryTTS and present examples that support the argument that this does not meet our requirements to avoid monotony in synthesized speech.

Different Voice Quality, Same Source of Errors

Communicative prosody is a different issue than voice quality and, consequently, should be studied and evaluated separately. The synthesizers and voices used for each technique, namely: concatenative or unit selection (US) and statistical, distinguishing with respect to the latter between Hidden Markov Models (HMM) and neural networks (NN), are detailed in Table 6.29. The chosen gender is female and the language American English in all speech synthesizers. As the focus of this dissertation is on the linguistic rather than on the signal processing component, I will not elaborate on the techniques. Further reference to the signal processing component and differences between synthesis techniques can be found in the literature; see, e.g., (Tabet and Boughazi, 2011; Watts et al., 2016).

Example (2)[11] shows the thematicity annotation of the sentence: *Rolls-Royce Motor Cars Inc. said it expects its U.S. sales to remain steady at about 1,200 cars in 1990.* The sentence has been chosen to exemplify that different types of synthesis techniques (i.e., US, HMM and NN) make similar errors in connection with the lack of communicative structure.

[9] http://mary.dfki.de/

[10] https://text-to-speech-demo.mybluemix.net/

[11] Sentence number 2 in our corpus.

Table 6.29: Synthesizers and voices used in the comparison of speech synthesis techniques.

Technique	Synthesizer	Voice name
US	MaryTTS	cmu-slt
HMM	MaryTTS	cmu-slt-hsmm
NN	Bluemix	Allison

(2)

[[*Rolls - Royce Motor Cars Inc.*]T1(SP1) [*said*]R1(SP1)]SP1 [*it*]T1 [*expects its U.S. sales to remain steady at about 1,200 cars in 1990*]R1 .

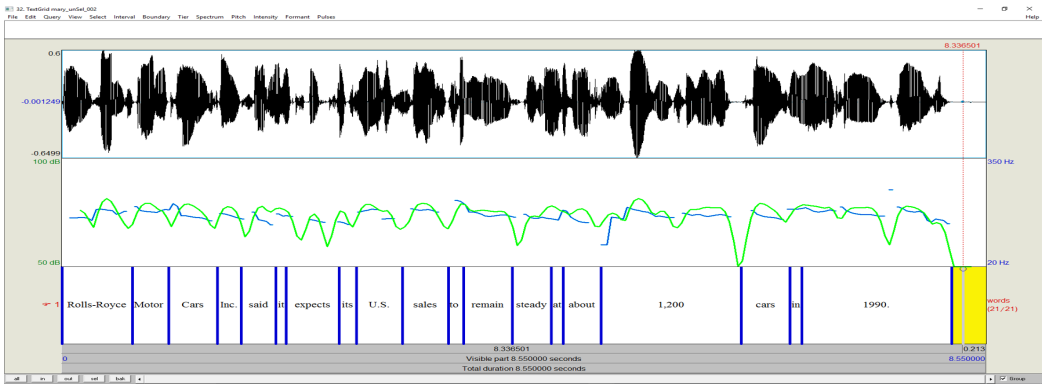


Figure 6.11: Unit Selection synthesis: Example (2) by MaryTTS.

Apart from the fact that both US and HMM voices by MaryTTS do not pronounce the abbreviation *Inc.* as *Incorporated*[12] while the NN by Bluemix does (and so do all our human speech samples in our corpus), the main issue affecting voice quality in these systems is connected to smooth transitions across syllables. The main audible difference between the US and the HMM and NN samples is the perceivable sound cuts between syllables due to the processing of concatenated units in the US voice. These cuts are even visually seen in Praat's representation of the F0 contour in Figure 6.11 for almost every syllable, whereas Figures 6.12 and 6.13 show more continuous F0 contours across syllables. This problem in the generation of smooth transitions across syllables in US voices has a direct impact when prosody modifications are tested. Thus, the resulting modified speech is

[12]The problem with the abbreviation *Inc.* in MaryTTS is not connected to prosody generation and could be solved introducing a dictionary entry for the normalizer module to map it to the word *Incorporated*.

perceived as highly distorted. Moreover, this distortion directly impacts prosody modifications because an F0 contour that was not generated from the beginning is impossible to be reconstructed in a post-processing stage.

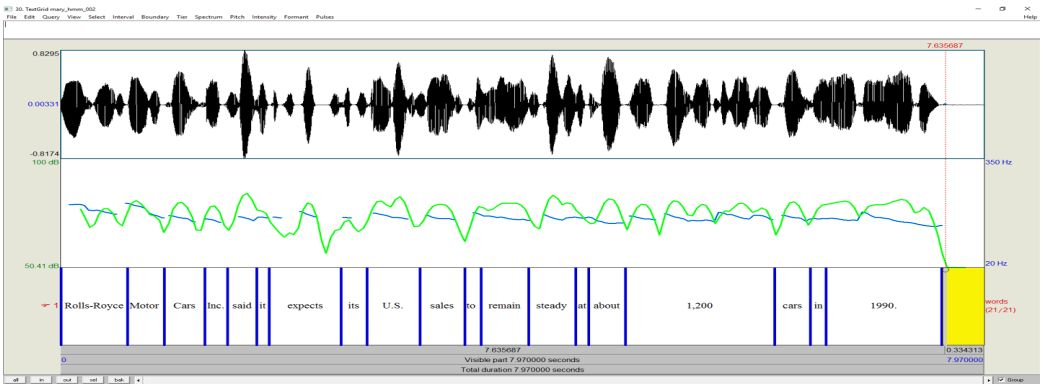


Figure 6.12: Hidden Markov Model synthesis: Example (8) by MaryTTS.

Concerning thematicity structure, (2) contains an initial specifier (SP1), *Rolls - Royce Motor Cars Inc. said*; a pronominal theme (T1), *it*, and a rheme (R1). *expects its U.S. sales to remain steady at about 1,200 cars in 1990*. Errors connected to the lack of communicative structure usually involve an inadequate placement of pauses between words. For instance, despite the higher voice quality of the NN sample compared to both US and HMM samples of (2), Figure 6.13 shows an awkward pause placement between *sales* and *to remain*. The fact that *to remain steady* is considered as a separate unit rather than as part of the same communicative span (in this case, the rheme R1) *said it expects its U.S. sales to remain steady at about 1,200 cars in 1990* is probably the reason behind the location of the pause that affects the perceived cohesion of information within the rheme span in the NN sample. However, neither US nor HMM samples include a break in this location.

A common characteristic in all three synthesized speech samples is the fact that none of them considers the specifier (SP1) *Rolls-Royce Motor Cars Inc. said* or any of its L2 partitions as segment that can be mapped to constitute a prosodic unit and, e.g., insert a pause after the embedded theme (T1(SP1)) *Rolls-Royce Motor Cars Inc.*. Instead, phonological rules apply for the generation of prosodic units in all words involved and, consequently, *Inc.* is connected to *said it expects* in all of the analyzed synthesized samples: US, HMM and NN.

Table 6.30 summarizes the performance of these synthesized speech samples

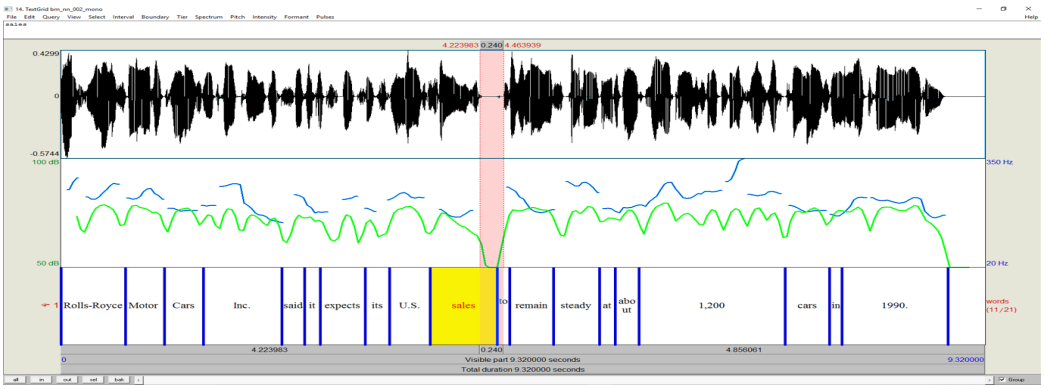


Figure 6.13: Neural Network synthesis: Example (8) by Bluemix.

Table 6.30: Performance of samples from US, HMM and NN speech synthesis.

	US	HMM	NN
Abbreviation	✗	✗	✓
Smooth transitions	✗	✓	✓
Cohesion within R1	✓	✓	✗
Break after SP1	✗	✗	✗
Break after T1(SP1)	✗	✗	✗

according to the aforementioned criteria. These criteria involve voice quality (at the upper part of the table) and the lack of communicative structure (at the lower part of the table). A tick means that the issue is correctly solved, and a cross means that it is not conveniently tackled. As can be observed, the NN sample makes errors in all three issues connected to communicative structure (namely, cohesion within the rheme span and break after the specifier or after the embedded theme). Even though these errors are compensated by the higher voice quality, the improvements proposed in this dissertation are argued to not only improve MaryTTS HMM voice, but also high quality commercial systems such as the NN voice by Bluemix. Next section exemplifies the main differences found in synthesized and human speech, following the assumption that a data-driven approach will contribute to the naturalness of synthesized speech.

Comparison of Synthesized and Human Speech Samples

In this subsection, human speech is compared to synthesized speech. As our objective is to generate more natural and expressive synthesized speech, the comparison between speech samples from our corpus and the synthesized output of

TTS applications is instrumental to understand the main challenges that have to be dealt with in the implementation of prosody enrichment. The statistical voices from HMM by MaryTTS and NN by Bluemix are compared to gold standard (human) speech samples from our corpus to underline the main differences in prosody by means of an example sentence. The ToBI representation is used to annotate speech samples and thus highlight the differences in prosodic phrasing and intonation contours.

Example (3)[13] shows the annotation of hierarchical thematicity of the sentence *The researchers said they have isolated a plant gene that prevents the production of pollen.* The sentence contains fifteen words, which is within the average length in our corpus. I analyze the commonalities across speech samples for this sentence.

(3)

[[*The researchers*]T1(SP1) [*said*]R1(SP1)]SP1 [*they*]T1 [*have isolated a plant gene* {[*that*]T1(P2) [*prevents the production of pollen*]R1(P2)}P2]R1.

Gold standard speech samples of (3) annotated with ToBI labels are displayed in Table 6.31. All human samples coincide in the segmentation of this sentence into two PPhs. The first PPh coincides with the beginning of the embedded proposition (P2) within the rheme span *that prevents the production of pollen*. Besides, the majority of samples contain a rising PA (L*+H) on *researchers*, which forms part of the embedded theme in the initial specifier span (T1(SP1)). All human samples also coincide in the falling ToBI pattern (LL%) at the end of the sentence, with some alternatives regarding where the previous PA is located: on either *prevent* or *production*. In some speech samples both words are prominent.

Figure 6.14 shows the Praat representation for one voice sample (by participant spk5f) of example (3) marking these segments. Listening to this sample and looking at the F0 contour lines (in blue from the Praat representation), there are three homogeneous prosodic contours. No BT is marked in PPh 1, as there is no actual break between *said* and *they* due to the "assimilation"[14] of two dental phonemes: /d/ and /ð/). The assimilation causes the F0 contour line to form a unified shape that has its own entity within the whole sentence. Thus, this segment forms a visible F0 contour (in the Praat representation depicted in Figure 6.14) that accounts for the perceived expressiveness of this human speech sample.

[13] Sentence number 62 in our corpus.

[14] The process of phoneme assimilation in the speech chain consists in the pronunciation of two phonemes as only one.

Table 6.31: Thematicity partition of example (3) and ToBI annotation of human speech samples.

	The researchers	said	they	have isolated	a plant gene	that	prevents	the production of pollen
L1	SP1		T1	R1				
L2prop						P2		
L2them	T1	R1				T1	R1	
spk1f	L*+H			L*+H	LL%		H*	LL%
spk1m	H*			H*	HL%		H*	LL%
spk2f	H*			H*	LL%		H*	H* LL%
spk2m	H*			L*	H*LL%		H*	L* LL%
spk3f	H*			L*	H*	LL%	L*	L* LL%
spk3m	L*+H			H*	L*LL%		L*	L* LL%
spk4f	L*+H			H*	H*HL%		H*	H*HL%
spk4m	L*+H			H*	HL%		H*	LL%
spk5f	L*+H			H*	LH%		H*	LL%
spk5m	L*+H			H*	H*		L*	LL%
spk6m	L*+H			L*+H	HL%		H*	LL%
spk7f	H*			H*	H*LL%		H*	LL%

However, looking at synthesized samples from HMM (see Figure 6.15) and NN (see Figure 6.16), F0 contours (also represented by the blue line in Praat) are much flatter in general, except for the PAs. The HMM sample by MaryTTS makes use mainly of lexical stress on content words to generate prosody, as it is made visible in the rather flat F0 contour shown in Figure 6.15. Intensity lines do not even group together *plant* and *gene*. Instead, *gene* is linked together with *that*; this connection is not found in human speech samples.

Even though the NN sample by Bluemix shows a higher variation in F0 prominence (apart from a better voice quality), it fails to mark a clear prosodic unit at the beginning of the embedded P2 as found in all human speech samples. This is even visible in the intensity contour that goes from the initial plosive sound (/p/) in *plant* to the next plosive in *prevents*[15]. This causes that, despite the fact that Bluemix links correctly the word compound *plant gene*, it is still not able to make the thematicity-based dissociation when the embedded P2 starts at *that*.

The brief analysis of this example reveals the fact that human samples from our corpus for this sentence consistently signal communicative spans by means of prosody, whereas the presented synthesized samples do not exploit this role

[15]Plosive phonemes are associated with a fall in intensity and F0 as they do not contain vibration of the vocal folds and involve a considerable amount of energy to make the actual implosion in human speech.

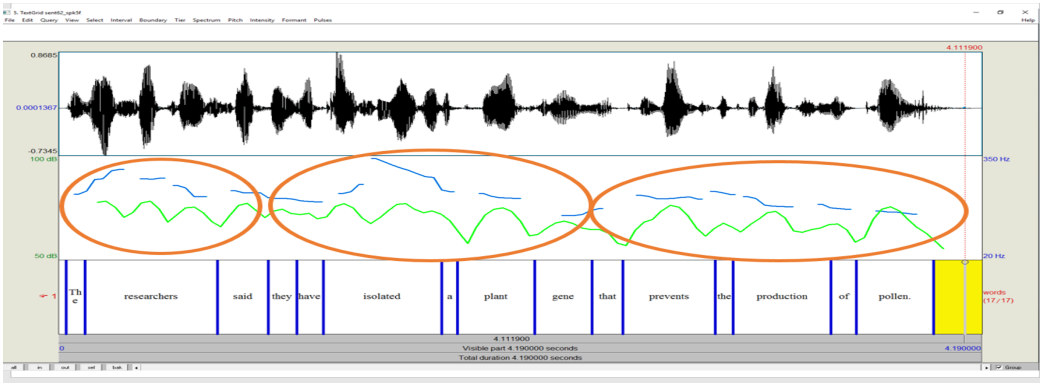


Figure 6.14: Human speech sample: Example (3) by spk5f.

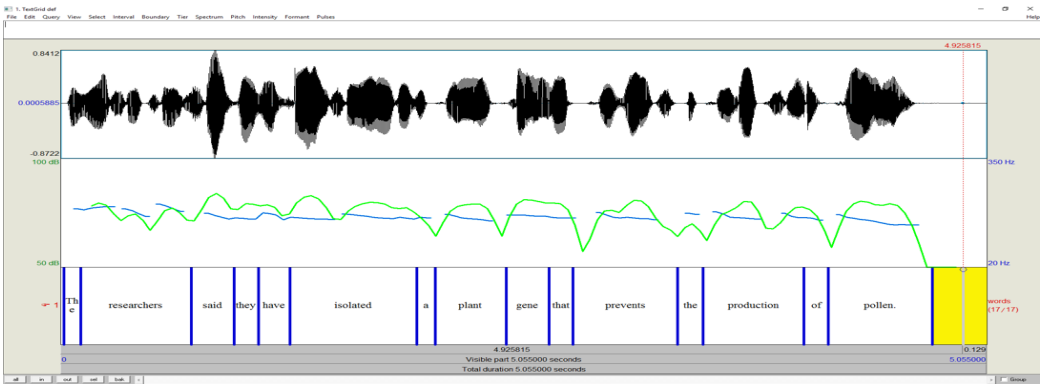


Figure 6.15: Hidden Markov Model synthesis: Example (9) by MaryTTS.

of prosody for generation of synthetic speech. Such differences are perceive in auditory environments, and can also be spotted in visual representations. In other words, there is a connection between perception and parametric representations that needs to be further explored.

Testing ToBI Labels in MaryTTS

Tests using ToBI labels have been made for prosody enrichment in MaryTTS. MaryTTS supports ToBI to force location and specification of type of accents and boundary tones within the MaryXML specification as explained in the documentation[16]. MaryTTS can be downloaded for local use and has a user interface shown in Figure 6.18. Selecting from the left window, the RAWMaryXML option a ready-to-use template is available. Once the text with the desired prosodic modification is inserted, the resulting modified speech can be played and downloaded

[16]http://mary.dfki.de/documentation/maryxml/index.html

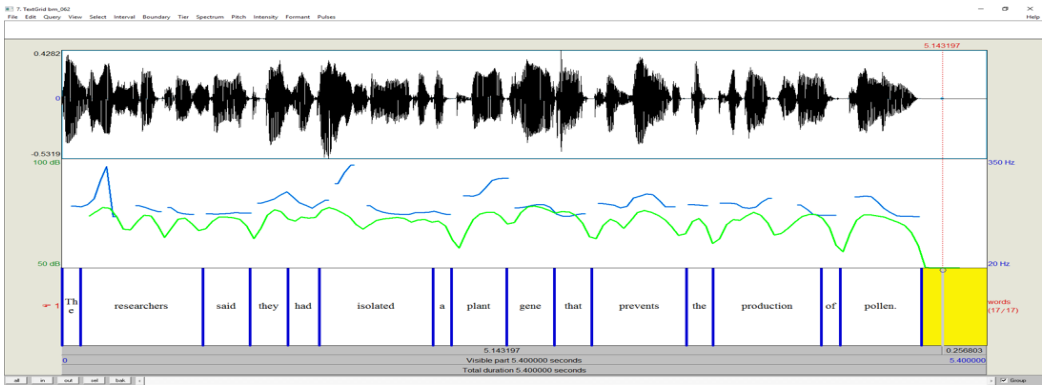


Figure 6.16: Neural Network synthesis: Example (9) by Bluemix.

clicking on the 'save' button. Figure 6.17 shows the collection of boundary tones defined in the MaryXML schema file accepted for German ToBI (g-ToBI) and English ToBI (e-ToBI) catalogs.

```
<xsd:simpleType name="gtobi_boundarytone.type">
    <xsd:restriction base="xsd:string">
        <xsd:enumeration value="H-"/>
        <xsd:enumeration value="h-"/>
        <xsd:enumeration value="!H-"/>
        <xsd:enumeration value="!h-"/>
        <xsd:enumeration value="L-"/>
        <xsd:enumeration value="l-"/>
        <xsd:enumeration value="H-%"/>
        <xsd:enumeration value="h-%"/>
        <xsd:enumeration value="!H-%"/>
        <xsd:enumeration value="!h-%"/>
        <xsd:enumeration value="H-^H%"/>
        <xsd:enumeration value="h-^h%"/>
        <xsd:enumeration value="!H-^H%"/>
        <xsd:enumeration value="!h-^h%"/>
        <xsd:enumeration value="L-H%"/>
        <xsd:enumeration value="l-h%"/>
        <xsd:enumeration value="L-%"/>
        <xsd:enumeration value="l-%"/>
        <xsd:enumeration value="unknown"/>
        <xsd:enumeration value="none"/>
    </xsd:restriction>
</xsd:simpleType>
<xsd:simpleType name="etobi_boundarytone.type">
    <xsd:restriction base="xsd:string">
        <xsd:enumeration value="L-L%"/>
        <xsd:enumeration value="l-l%"/>
        <xsd:enumeration value="H-H%"/>
        <xsd:enumeration value="h-h%"/>
        <xsd:enumeration value="H-L%"/>
        <xsd:enumeration value="h-l%"/>
    </xsd:restriction>
```

Figure 6.17: MaryXML schema file for ToBI boundary tones.

Example (4)[17] shows the thematicity segmentation of the sentence *The proposed changes also would allow executives to report exercises of options later*

[17]Sentence number 62 in our corpus.

and less often. Example (4) is used to test a rising boundary tone (LH%) inserted at the end of the theme span (T1) after the word *changes*.

(4)

> [*The proposed changes*]T1 [*also would allow executives to report exercises of options later and less often*]R1 .

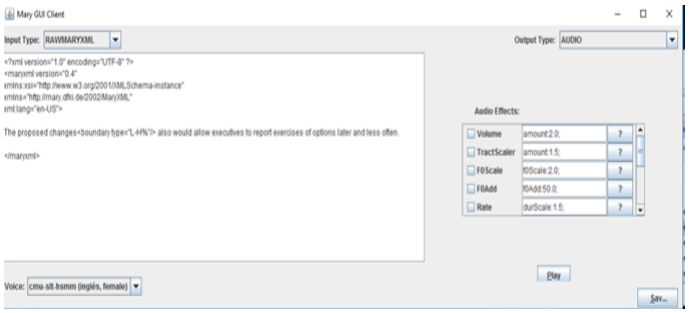


Figure 6.18: RAWMaryXML of example (4) in MaryTTS GUI.

The default output by MaryTTS without any prosody control tag contains, in fact, a very subtle rising of the F0 in the final syllable of *changes* (see Figure 6.19). Such a rising intonation is motivated by the implementation of *givenness* in the module (as described in Chapter 3).

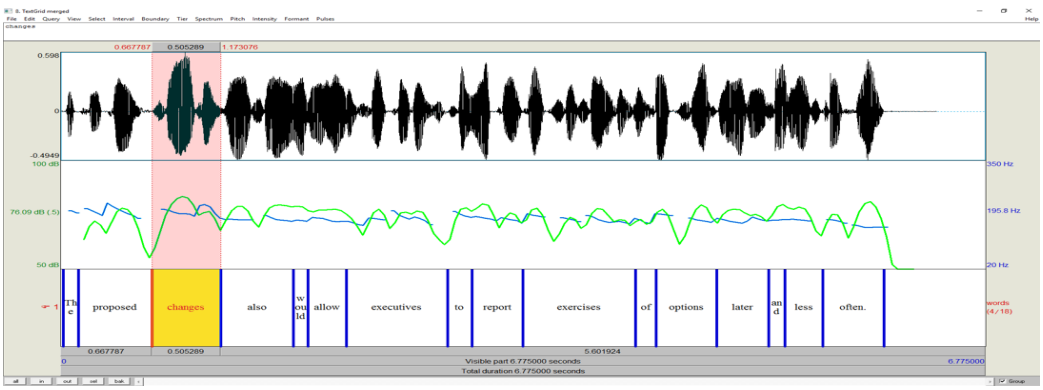


Figure 6.19: Default output of (4) by MaryTTS.

However, this rising intonation is hardly perceived as there is no break between the end of T1 and the beginning of R1. Two tests have been made to solve these two deficiencies:

1. Insertion of a rising boundary tone (LH%): the control tag <boundary type="L-H%"/>is inserted resulting in a break insertion and a falling tone, as Figure 6.20 shows.

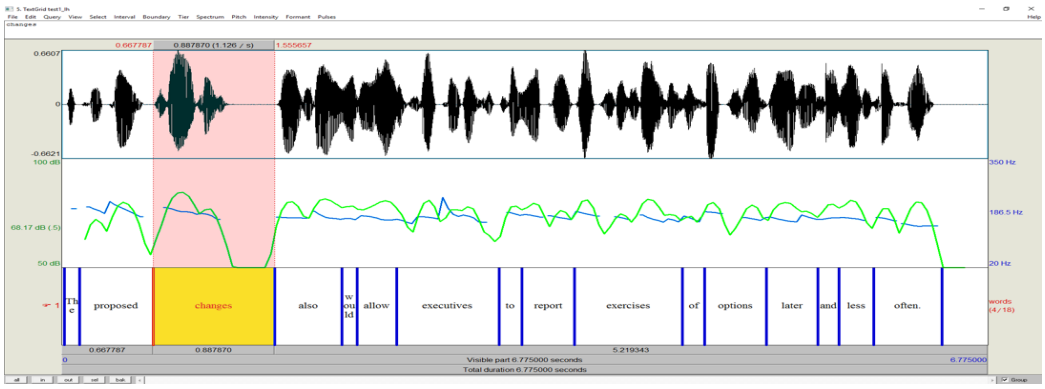


Figure 6.20: Output of (4) with boundary type modification by MaryTTS.

2. Insertion of a break without tone specification: the control tag <boundary duration="100"/>is inserted, resulting in a break insertion and the falling tone modification as shown by the pitch line in Figure 6.21.

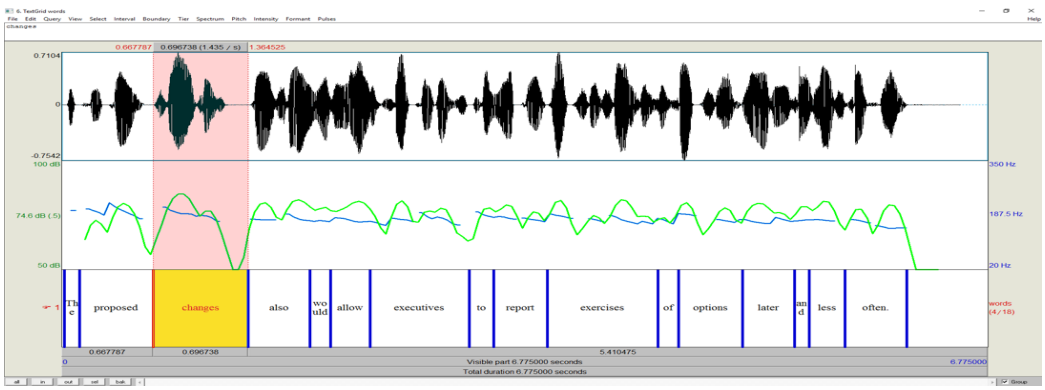


Figure 6.21: Output of (4) with boundary duration modification by MaryTTS.

After further testing with different sentences and types of prosodic modifications using ToBI labels, some conclusions can be drawn:

- the most relevant ToBI labels for the implementation of rising tunes in themes (among other spans) are not fully implemented in MaryTTS, in particular the L-H% boundary tone;

- boundary control tags may cause unexpected modifications beyond the actual specification that is introduced; for instance, in example (4), the introduction of a break caused the subtle rising contour (associated to the concept of givenness in MaryTTS) at the end of T1 disappear and a falling tune was forced together with the missing break that was pretended;
- prosody control tags are voice and language dependent, which makes ToBI not scalable for testing in other voices, languages, and even TTS applications.

6.5.2. Thematicity-based Prosody Enrichment

The thematicity-based prosody enrichment presented in this section departs from a parametric representation of prosody, as ToBI labels were proved to yield suboptimal results, especially for the control of rising intonation patterns (LH%), which are essential (as already demonstrated) in the thematicity–prosody correspondence. This does not mean that ToBI should be left aside in the implementation of the information structure–prosody interface. However, mapping ToBI labels to the actual prosodic modifications is a task on its own, and further research efforts are needed that are out of the scope of this dissertation.

As previously shown, a tripartite hierarchical thematicity provides a better correspondence with prosody in classification experiments, and proves to be more appropriate than a binary flat representation (see Section 6.2). In this section, experiments are presented on the use of a tripartite hierarchical thematicity in the context of implementing prosody modifications in a CTS application. In particular, the analysis of the distribution of three normalized prosodic parameters ('z_F0', 'z_int' and 'z_sr') is used to implement prosody modifications of overall values along broad text segments that coincide with thematicity spans.

The implementation of the prosody enrichment is carried out using MaryTTS (Schröder and Trouvain, 2003). Several reasons support this choice:

- it is open source;
- it supports SSML tags and ToBI labels;
- it is well documented and fully maintained;
- it allows the creation of new voices;
- it has an API and web interface for both online and local testing.

The default MaryTTS input text has been enriched using MaryXML prosody specifications.[18]

[18] https://github.com/marytts/marytts/wiki/ProsodySpecificationSupport/ and

Experimental Setup

Table 6.32 displays the thematicity partitions that are chosen for the prosody enrichment perception test, the sentence number in this section ('Ex.') and corresponding reference in our corpus ('Co.'). In the case of embedded propositions, monosyllabic embedded spans have not been taken into account in this experiment, since prosodic modification would be otherwise dependent on phonological and syntactic structures of the sentences rather than on the communicative segmentation. The selected sentences include the following thematicity structures:

- Binary division: theme (T1) rheme (R1) in examples 1 and 2;
- Tripartite division: specifier (SP1) theme (T1) theme (R1): with SP1 in initial and final positions in examples 4 and 6 respectively;
- Embedded Proposition (P2): in theme (examples 3, 5 and 6) and rheme spans (P2(T1) and P2(R1) respectively) in examples 4.

Table 6.32: Selected sentences for perception test annotated with thematicity.

Ex.	Co.	Thematicity partition
1	3	[*The luxury auto maker*]T1 [*last year sold 1,214 cars in the U.S.*]R1
2	20	[*For its employees to sign up for the options*]T1 , [*a college also must approve the plan.*]R1
3	57	[*Mr. Mayor's hope that*]T1 [*references to "press freedom" would survive unamended*]P2(T1) [*seems doomed to failure*]R1
4	62	[*The researchers said*]SP1 [*they have isolated a plant gene*] [*that prevents the production of pollen.*]P2(R1)
5	91	[*When he sent letters* [*offering 1,250 retired major leaguers*]T1(P3) [*the chance of another season*]R1(P3)]T1 , [*730 responded.*]R1
6	96	[*Men who have played hard all their lives*]T1 [*aren't about to change their habits*]R1 , [*he says.*]SP1

Prosody Representation Prosodic acoustic parameters were automatically extracted using the extension of Praat for feature annotation introduced in Chapter 5. These parameters include mean and standard deviation of F0, intensity and speech rate. Then, for each themacity span, mean acoustic parameters of F0, intensity and speech rate were normalized to z-scores relative to the whole sentence.

Each thematicity span is annotated with z-scores of three prosodic parameters relative to the corresponding sentence: z-score of F0 (z_F0), intensity (z_int) and

http://mary.dfki.de/documentation/maryxml/index.html. These specifications are based on the *Speech Synthesis Markup Language* (SSML) (Taylor and Isard, 1997) recommendation https://www.w3.org/TR/speech-synthesis/

speech rate (z_sr). Figure 6.22 shows this segmentation in the *Praat on the Web* platform[19] and the automatically annotated prosodic parameters for T1 at L1 for one speech sample from the corpus. Z-scores represent positive and negative deviations relative to the sentence mean value of each parameter. Such deviations are used to analyze whether different parametric distributions occur between hierarchical thematicity spans using the average values across speakers in our corpus.

Figure 6.22: Example of hierarchical thematicity and annotation of prosodic parameters.

All speech samples are segmented according to the thematicity annotation, and prosodic parameters are computed for each segment. Spans are grouped by thematicity labels and prosodic parameters are averaged across speakers. Then, the distribution of normalized prosodic parameters according to hierarchical thematicity is analyzed across samples of read speech from twelve participants. Finally, a selection of thematicity spans is done and their parametric distribution extracted from the corpus analysis is used to derive prosody modifications for a TTS application.

Parameters Distribution in Hierarchical Thematicity The analysis of the distribution of average z_int, z_F0 and z_sr is presented in several figures, grouped in level 1 (L1) (cf., Table 6.33) and level 2 (L2) thematicity (cf., Table 6.34).

Table 6.33 shows that there is a distinct distribution of parameters between T1, R1 and SP1 spans within the L1 thematicity spans. T1 displays positive deviations (highlighted in bold) in all z_int, z_F0 and z_sr parameters; R1 has negative z_int

[19]http://kristina.taln.upf.edu/praatweb/

Table 6.33: Distribution of prosodic parameters in L1 thematicity.

	z_int	z_F0	z_sr
T1	**0.16**	**0.47**	**0.95**
R1	-0.04	-0.17	**0.30**
SP1	-0.07	**0.15**	**1.76**

and z_F0 and positive (but lower than T1) z_sr; and SP1 is characterized by negative z_int (like R1), positive (but lower than T1) z_F0 and positive (higher than T1) z_sr. These figures suggest that each L1 thematicity partition has its own distinct prosodic characteristics. These diverse prosodic characteristics indicate, for instance, that themes are pronounced louder, with a higher overall F0 and faster speech rate than rhemes and with a similarly fast speech rate with respect to specifiers.

Table 6.34: Distribution of prosodic parameters in L2 thematicity.

	z_int	z_F0	z_sr		z_int	z_F0	z_sr		z_int	z_F0	z_sr
T1(T1)	**0.27**	**0.57**	**0.23**	R1(T1)	**0.27**	**0.80**	**0.89**				
T1(R1)	**0.12**	**0.21**	**1.30**	R1(R1)	-0.23	-0.41	**0.48**				
T1(SP1)	-0.13	-0.07	**2.25**	R1(SP1)	-0.21	-0.21	**1.11**	SP1(SP1)	**0.25**	-0.06	**3.13**

Table 6.34 shows the average normalized parameters for embedded thematicity spans in three sections for embedded themes, rhemes and specifiers respectively. The level 1 spans where level 2 thematicity is embedded are represented in rows. Thus, level 2 specifiers are embedded only in a level 1 specifier, i.e., SP1(SP1), in the corpus, even though it is possible to find level 2 specifiers in any other span.

Embedded themes (left column in Table 6.34) in either theme, i.e., T1(T1), and rheme, i.e., T1(R1), spans show positive deviations in all parameters. T1(T1) shows higher values in z_int and z_F0 and lower z_sr than T1(R1). Embedded rhemes (central column of Table 6.34) share a negative tendency in intensity and F0 values and a positive value of speech rate, if they are embedded in either rheme R1(R1) and specifier R1(SP1) spans. However, rhemes embedded in a theme span, i.e., R1(T1), show a similar parametric distribution as themes that are embedded in a theme span, i.e. T1(T1), as both share positive values in all acoustic parameters and even the same value for z_int (0.27). However, R1(T1) shows

higher z_F0 (0.80) and z_sr (0.89) than T1(T1).

Summing up, the analysis of average z_scores across speakers shows a distinct distribution pattern of prosodic parameters extracted from different thematicity spans. If we compare these results to previous studies on the information structure–prosody interface that related themes with rising and rhemes with falling F0 contours, the distribution of z_F0 values across our corpus also supports the argument that themes involve a higher z_F0. This argument is further extended to the concept of embeddedness: spans that are embedded in themes (both T1(T1) and R1(T1)) have higher values of z_F0 than other embedded spans. Differences in the distribution of values observed in thematicity elements will be exploited as a data-driven approach for the derivation of a thematicity-based generation of SSML prosody control tags.

Testing the information structure–prosody correspondence in TTS

The proof of concept presented in this section consists in an automatic transformation of sentences annotated with thematicity into an SSML format output for prosody enrichment of synthesized speech. As already mentioned, the SSML *prosody tag* takes six optional attributes (overall pitch, pitch contour, pitch range, speech rate, duration and volume). Three of them, namely, overall pitch, speech rate and volume with relative values are chosen for the demonstration. Other attributes were also tested, but they did not yield noticeable changes so they were discarded for this experiment.

Table 6.35: Conversion of acoustic parameters to SSML attribute values.

Thematicity	z_int	'volume'	z_F0	'pitch'	z_sr	'rate'
T1	0.15	15	0.50	50	0.95	35
R1	-0.05	-5	-0.20	-20	0.30	10
SP1	-0.10	-10	0.15	15	1.00	35
R1(T1)	0.30	30	0.50	25	0.25	15
R1(R1)	-0.25	-25	-0.40	-15	0.50	25

A set of examples from the corpus is selected for the assessment of the thematicity-based prosody enrichment in MaryTTS using a statistical voice. The distribution of prosodic parameters in the whole corpus (described in the previous section) is mapped onto the values that each attribute of the SSML tag will take. Initial testing of SSML attributes proved that the most convenient prosody modifications

were achieved when only one attribute was modified along the same sentence, varying the values according to the thematicity span. Table 6.35 presents the characterization of thematicity for the spans selected for the implementation.[20] Some values (especially those for speech rate) were scaled to an appropriate percentage, because previous testing using SSML prosody tags showed an undesired distortion when a very high attribute value was inserted. For instance, if an increase of 95% in speech rate was specified, the resulting speech would sound far too fast with an associated F0 increase, and consequently unnatural and, sometimes, unintelligible.

The overall pipeline within a CTS application is described in Figure 6.23. As can be seen, the thematicity-based prosody module takes as input an annotated file with thematicity in *txt* or *CONLL* format to begin the processing. The core part of the tool is a pair list of attributes and thematicity spans with their corresponding values. This parameter–span list of values is built using a data-driven approach on a corpus of read speech, annotated with thematicity. In the first place, the module splits text into sentences and assigns each sentence a prosody attribute (either pitch, volume or speech rate). This attribute is varied in each sentence so as to allow a range of prosodic variability achieved by means of different prosodic elements. This variability is aimed at breaking the monotony of the synthesized speech. Then, a query to the pair list that contains percentages for each thematicity span[21] is performed to assign the value of the selected attribute. In order to attain more variability, the final value of the prosody attribute is randomly assigned within a range of $\pm 5\%$ from the pair list entry value for each span.

6.5.3. Evaluation

The evaluation of a selection of sentences with thematicity-based prosody modification is done following two strategies: perception tests and similarity to gold standard. This combination of subjective and objective criteria is expected to provide not only an assessment of the actual changes, but also insights into what needs to be further improved in the implementation of prosody generation in synthesized speech.

Perception Tests

Thirty synthesized speech samples have been evaluated in a perception test taken by thirty participants. Thematicity-based prosody modifications were com-

[20] Figures are round up to the closest half tenth.

[21] Percentages are computed from the mean z-score values for each thematicity span in our corpus.

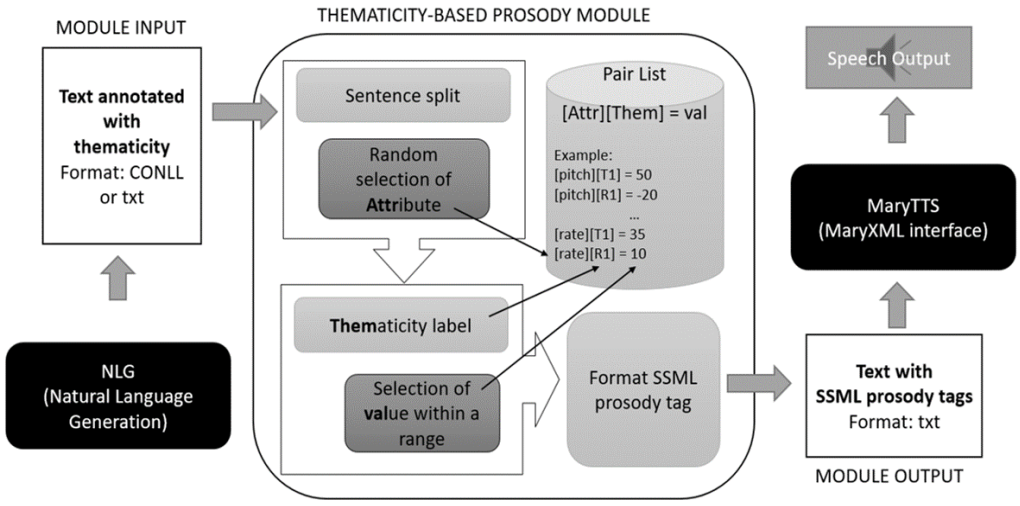


Figure 6.23: Thematicity-based prosody enrichment pipeline within a CTS application.

pared to the default output by MaryTTS (which is the baseline for comparison). Modifications included F0, speech rate and breaks corresponding to thematicity spans in isolation and in combination. Intensity was excluded from the evaluation as there was no perceivable change in the modified synthesized sentence. The perception test consisted in two parts: (i) a *Mean Opinion Score* (MOS) test rating within a Likert scale from 1 to 5 at the level of expressiveness[22] and; (ii) a pairwise comparison where the most expressive sentence is chosen between the following pairs: baseline–F0, combination–break and F0–combination. A total amount of 1,440 answers are considered in the evaluation.

Table 6.36: Evaluation: MOS test results.

	baseline	F0	speech rate	break	combination
sent 1	2.03	**2.67**	2.30	**2.53**	2.43
sent 2	2.97	3.07	3.10	2.83	3.00
sent 3	2.93	2.83	*2.40*	*2.37*	2.90
sent 4	2.83	2.67	2.90	2.70	2.60
sent 5	3.03	2.87	2.73	3.00	3.06
sent 6	2.40	2.67	**3.17**	2.73	2.17
Average	2.70	2.79	2.77	2.69	2.61

[22]Defined as *effectively conveying meaning*.

Results from the MOS test are depicted in Table 6.36. T-tests were performed to observe the level of significance with a confidence of 95%. Bold figures represent statistically significant improvements over the baseline (t-test, p $<$0.05), and figures in italic represent those results in which the lower value is also statistically significant (t-test, p $<$0.05) with respect to the baseline. F0 and speech rate modifications are rated higher than the default voice. Looking at each specific sentence, thematicity-based modifications tend to be rated higher. Sentences 1 and 6 show statistically significant differences with respect to the baseline for F0 and break modifications (in sentence 1) and for speech rate in sentence 6. Sentence 3 shows statistically significant worse results for modifications concerning speech rate and break compared to the baseline.

Results from the pairwise comparison are reported in Table 6.37 as the average of selected answers from 0 to 1. In this case, F0 and break modifications are preferred over both the baseline and combination of prosodic modifications.

Table 6.37: Evaluation: pairwise results.

	baseline	F0	combined	break	F0	combination
sent 1	0.20	**0.80**	0.50	0.50	**0.77**	0.23
sent 2	0.50	0.50	0.43	**0.57**	**0.77**	0.23
sent 3	0.50	0.50	0.37	**0.63**	0.40	**0.60**
sent 4	0.47	**0.53**	0.30	**0.70**	0.47	**0.53**
sent 5	**0.60**	0.40	0.47	**0.53**	0.33	**0.67**
sent 6	0.50	0.50	0.40	**0.60**	**0.70**	0.30
Average	0.46	**0.54**	0.41	**0.59**	**0.57**	0.43

Let us analyze the sentences of the evaluation in detail to take a closer insight on the results:

1. *The luxury auto maker last year sold 1,214 cars in the U.S.* The modification of F0 is the best ranked modification in both MOS and pairwise comparison. A higher pitch modification in the theme span *The luxury auto maker* is remarkably perceived as more expressive. The insertion of a break after the theme span also shows statistically significant better results in the MOS test. In this sentence, the theme contains four words and the sentence has a total of eighteen (counting the numeral *1,214* as six words, that is, *one thousand two hundred and fourteen*). In this eighteen word sentence, no break is inserted in the default sentence by MaryTTS. This may be the reason why both the break or the contrast in F0 modifying the theme span yields better results in perception.

2. *For its employees to sign up for the options, a college also must approve the plan.* The default synthesis of this sentence poses a problem that is difficult to overcome in a post-processing prosody modification. In other words, the comma forces the TTS system to impose a specific prosody that cannot be easily overridden by SSML prosody tags. This predetermined prosody involves not only a break, but also a sharp change in F0 after the comma. This may be the reason why modifications of F0, speech rate and the combined tag achieve slightly, but not statistically significant results in the MOS test and similar results in the pairwise comparison.

3. *Mr. Mayor's hope that references to "press freedom" would survive unamended seems doomed to failure.* In this sentence, the default synthesis includes a break after *"press freedom"*, which is bound to be motivated by the double quotes. In this case, the SSML prosody tags that modify the speech rate (that is the rate and the combined tag) succeed in overriding the inadequately placed break and inserting a more natural break that coincides with the beginning of the rheme span, i.e., *seems doomed to failure.* Surprisingly enough, the modification of the rate and insertion of break corresponding to thematicity spans achieve statistically significant worse results in the MOS test compared to the baseline. The reason might be the ambiguity in the word *hope*, which can be a verb or a noun, and it is not until you read through the whole sentence that you may realize it correct function. This ambiguity might have misled participants in the test, which could be the reason for the bad results in perception tests of the prosody modifications.

4. *The researchers said they have isolated a plant gene that prevents the production of pollen.* Results of the MOS test indicate that only the modification of speech rate is considered slightly better than the default. This modification specifically makes that the embedded propositional content within the rheme span *that prevents the production of pollen* is spoken at a slower speech rate. The sentence contains a high syntactic complexity that is also reflected in the communicative structure. However, due to the fact that the default synthesis already includes a rising tone at the end of the word *gene*, i.e., right before the R1(R1), the performed modifications do not add any improvement. In this case, the communicative span coincides with a syntactic boundary that is taken into account by the speech synthesizer to generate a suitable F0 contour. This indicates the importance of considering high-level linguistic information in the generation of prosody.

5. *When he sent letters offering 1,250 retired major leaguers the chance of another season, 730 responded.* This twenty-four word-long sentence (including numerals) has a complex syntactic and thematicity structure. The

thematicity annotation involves a long propositional theme span containing an embedded proposition with another long theme and rheme:

(5)

[{[When]SP1(P2) [he]T1(P2) [sent letters {[offering 1,250 retired major leaguers]T1(P4) [the chance of another season,]R1(P4)}P4]R1(P2)}P2]T1 [{[730]T1(P3) [responded.]R1(P3)}P3]R1

The default synthesis, as previously remarked for sentence 2, imposes some restrictions over prosody modifications due to the punctuation mark and its predetermined break and sharp decrease in F0 after the comma. This might lead to the modestly better result in both MOS and pairwise test of the combined prosody modification. This modification includes a pause and audible variation in speech rate and F0 in the embedded rheme span R1(P4) *the chance of another season*, which is instrumental for the understanding of this rather long and complex utterance.

6. *Men who have played hard all their lives aren't about to change their habits, he says.* The modification of speech rate in this final example has yielded statistically significant better results than the baseline in the MOS test. This modification consisted in lowering the speech rate considerably in the embedded proposition within the theme span *who have played hard all their lives*. In this case, the previously undesired effect of lowering F0 after the comma, does not affect the output, as in fact, this is the intended effect when the thematicity–prosody correspondence is considered in specifiers for reported speech (e.g., *he says*).

Summing up, F0, break and speech rate enrichments based on thematicity spans are perceived, in general, as more expressive than the baseline. This contributes to the idea that communicative spans are important in generating expressive synthesized speech, and that a variety of prosodic cues contributes to signaling the information structure–prosody correspondence. However, the inadequate prosody rendering of the default synthesized output often linked to punctuation marks often impedes that the expected prosody modification is conveniently produced. Therefore, statistically significant results are seldom obtained when comparing the prosody enriched sentences to the baseline.

Modifications of speech rate perform better than initially expected in the perception of thematicity-based prosody enrichment. This good performance of speech rate supports the argument that prosody variation goes beyond F0 contours and break insertion.

Finally, it must be mentioned that participants who took part in the perception test remarked, in general, that in most of the sentences, prosody modifications were barely perceivable and resembled too much to each other, such they found it difficult to decide which one was better than the rest. In what follows, the evaluation of these sentences is addressed by means of objective metrics that measure the distance of the synthesized sentences to a gold standard.

Objective Evaluation

The aim of the objective evaluation is to compare synthesized and human speech samples using acoustic parameters. Thus, by means of objective metrics the distance between synthesized samples to the gold standard is computed, so as the closer the distance to the gold, the more similar the synthetic voice. Table 6.38 presents the parameters included in this evaluation, segments used for extraction of values, and the metric applied to them in order to compute the distance between synthetic and gold samples. F0 and intensity are normalized as the distance to the mean in semitones and dB respectively.

Table 6.38: Objective evaluation: acoustic parameters.

Aspect	Parameter	Segment	Metric
general	intensity	sentence	KLD
	F0	sentence	KLD
prominence	number of PA	sentence	ED
	intensity	intensity peaks	KLD
	intensity	syllable nuclei (grouped in S, 2s, u)	KLD
	F0	syllable nuclei (grouped in S, 2s, u)	KLD
phrasing	number of BT	sentence	ED
	intensity	intensity valleys	KLD
speech rate	duration (in sec)	sentence	ED
	speech rate (words per sec)	sentence	ED
	speech rate (syllable nuclei per sec)	sentence	ED

Automatically detected peaks have been manually labeled as carrying primary stress (S), secondary stress (2s), and unstressed syllables (u). Samples have also been segmented into prosodic phrases, and prominence has been marked following the guidelines introduced in Chapter 4. Euclidean distance (ED) is computed for duration, speech rate, number of pitch accents (PA), and boundary tones (BT) at the sentence level according to Equation 6.1. For those parameters that involve

a Gaussian distribution, KLD is computed according to Equation 6.2.

$$ED(p,q) = \sqrt{(\sigma_p - \sigma_q)^2} \tag{6.1}$$

$$KLD(p,q) = log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \tag{6.2}$$

Finally, distance scores from all parameters are averaged to obtain an overall distance score for each sentence sample. Table 6.39 summarizes the results on the overall distance to a gold standard female human voice ('spk5f' in our corpus) for the baseline and the highest rated modification in the MOS test.

Table 6.39: Objective evaluation: distance scores.

samples	sent 1	sent 2	sent 3	sent 4	sent 5	sent 6
baseline	2.65	22.73	1.15	21.60	2.21	4.29
modification	3.01	9.95	0.83	38.43	2.17	2.68
difference	-0.36	**12.78**	**0.32**	-16.83	**0.04**	**1.61**

The difference (subtracting baseline from modification scores) shows positive results, and thus, a closer distance to gold for the modifications in all sentences, except for sentences 1 and 4. Let us now analyze the gold standard and synthesized sentences from a closer perspective to gain insight into these results. As the synthesized samples have already been discussed in the subjective evaluation, the focus now will be put on the human samples.

1. *The luxury auto maker last year sold 1,214 cars in the U.S.* The most outstanding characteristic when listening to the gold sample is the expressiveness achieved with a varied range of F0 contours and prominent words compared to both synthesized samples. In quantitative terms, the theme span in the gold sample has a higher average F0 than the rheme, which coincides with the prosody modification that achieves the highest score in the perception test. However, as the objective evaluation takes into account F0, intensity and duration parameters of unstressed, stressed and secondary stressed syllables, the modification gets slightly worse results caused by an associated increase in intensity scores when the F0 modification is applied, which results in a bigger distance to the gold sample than the default synthesis.

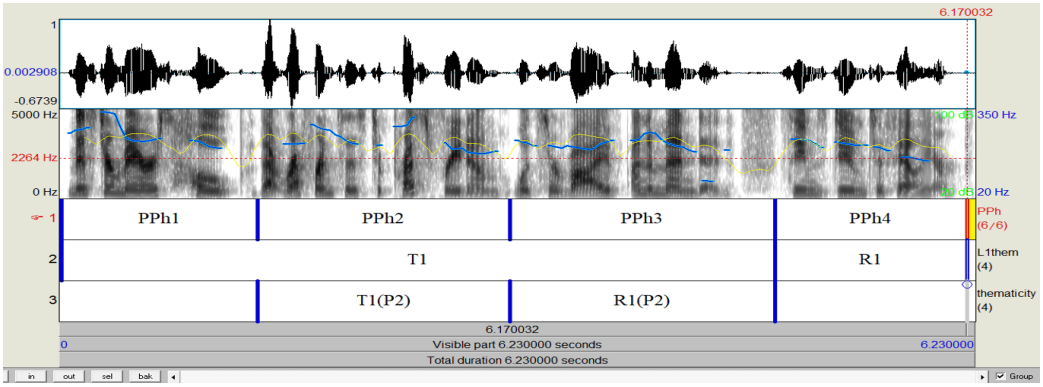


Figure 6.24: Coincidence of PPh and thematicity spans in sentence 3 by spk5f.

2. *For its employees to sign up for the options, a college also must approve the plan.* The synthesized samples presented serious deficiencies, as already mentioned, that could not be dealt with in the prosody enrichment. Consequently, there is a considerable objective distance in both samples. The modification gets closer to the gold especially due to the fact that intensity of valleys is indirectly affected when applying the speech rate modification.

3. *Mr. Mayor's hope that references to "press freedom" would survive unamended seems doomed to failure.* The most relevant difference between the gold sample and the synthesis is that this speaker clearly marks phrases, that is, she includes a break that coincides with thematicity spans, as shown in Figure 6.24. This is the reason why the prosody modification combining break, F0 and speech rate achieves closer scores to the gold sample as well as in the perception test.

4. *The researchers said they have isolated a plant gene that prevents the production of pollen.* This sentence clearly shows that the gold sample makes a contrast, especially in the rheme span *that prevents the production of pollen* using a lower speech rate. This coincides with the modification of speech rate attaining attaining the best rate in perception tests. However, the modification performs poorly in objective scores, especially those measuring F0 values in comparison with the gold sample when applying the speech rate modification.

5. *When he sent letters offering 1,250 retired major leaguers the chance of another season, 730 responded.* The gold sample for this sentence displays a combination of prosodic cues involving breaks, F0 and speech rate, and so does the best ranked synthesis modification in the perception tests. As already mentioned, this twenty-four word sentence has a complex thematicity

structure. The gold sample contains breaks at several places that coincide with thematicity spans, i.e., before the embedded rheme *the change of another season* and the main rheme *730 responded*. In the last case, the break is signaled in the text by a comma, which is reproduced in the default synthesis. However, as already mentioned, the sharp decrease in F0 of the default after the comma could not be fully overridden in the prosody modification. This might be the reason why in the objective score, the combined prosody modification performs slightly better than the default.

6. *Men who have played hard all their lives aren't about to change their habits, he says.* The most outstanding difference in this sentence between the gold sample and the default is that there is a break that coincides with the beginning of the rheme span *aren't about to change their habits* in the gold sample, which does not appear in the default synthesis. The prosody modification varying the speech rate in the preceding span achieves a contrast that is closer to the gold sample even though there is no actual pause between the spans.

When results from the objective evaluation and the perception tests are analyzed together, it can be observed that sentences including thematicity-based prosody modifications that achieve the highest score in the perception tests are those that involve a combination of acoustic parameters, speech rate and F0. However, these prosody modification do not vary greatly the default synthesis as they are done as a post-processing of already generated prosody, and in some cases, inherited characteristics cannot be changed by means of automatically derived SSML prosody tags. On the other hand, there is still a lot of work to be done in the computation of objective metrics to compare gold standard to synthesized samples in the speech community. It is not uncommon to read publications that only include subjective evaluations of the speech synthesis. And even though this was not one of the main objectives of this dissertation, it was attempted to establish some criteria that could serve for further investigation of how to assess expressiveness and naturalness in synthetic speech prosody.

When audibly comparing gold standard to synthetic samples, it is quite obvious that from the perspective of prosody generation, there is still a long way to reach a reasonable level of expressiveness and naturalness in synthetic speech. There are many aspects that are not being considered in this implementation experiment, which stand out when listening to synthetic and gold samples. For instance, the generation of different types of prominence depending on the location of the prominent word within a specific thematicity span. However, taking into account the limitations of the experiment presented in this section, both subjective and objective metrics demonstrate that a range of prosodic cues derived from

thematicity spans contribute to a more expressive and closer to natural speech prosody generation.

6.5.4. Discussion

Given the relevant role of the information structure–prosody interface in human communication, it seems reasonable that next generation virtual assistants face new challenges in adopting communicatively-oriented models. Current speech technologies have been oblivious to advances in theoretical fields studying this correlation, basically due to the lack of a formal representation of the communicative (or information) structure and limited representation of prosody to achieve variability in implementation settings.

As already mentioned in previous chapters, studies on human speech show that the understanding of a sentence improves when information structure is signaled by prosodic cues. In speech synthesis, however, understanding a message has traditionally been ascribed to the concept of *intelligibility*, and, thus, research has striven for improving signal processing techniques related mainly to voice quality, leaving aside communicative aspects related to how prosody helps to structure the content of a message.

However, in this section I argue that even good quality commercial synthesized voices do not account for communicative structure to generate prosody, and subtle changes (when they are communicatively derived) are perceived as more expressive. Moreover, the implementation of a thematicity-based prosody enrichment contributes in several aspects to the state of the art: (i) a formal description of communicative structure is used; (ii) hierarchical thematicity is annotated following established guidelines; (iii) prosodic representation is automatically computed and, consequently, time-consuming manual ToBI annotation tasks are avoided; (iv) a derivation of prosody enrichment is done empirically, from a corpus of read speech.

All in all, the implementation introduced in this section pivots the transition from theoretical work on the information structure–prosody interface to the integration of a data-driven prosody enrichment to achieve more communicative synthesized speech. Such a methodology uses a formal description for the annotation of hierarchical thematicity and a representation of prosody based upon automatically computed acoustic parameters. Results in classification experiments, perception tests and objective metrics yield an improvement of the proposed methodology over standard techniques.

One limitation of the current study is that it only considers relative acoustic parameters over rather large text segments. Key aspects of prosody modeling, like F0 contour generation in terms of prominence and phrasing have not been taken into account. This may explain why the MOS test resulted in a score of '3' on the 5-value Likert scale and few modifications were considered significantly better than the baseline. Therefore, further research in this direction should be encouraged.

Chapter 7

CONCLUSIONS AND FUTURE WORK

> ”Whether the communication is written or verbal, formal or informal, the question must be asked as to whether or not it was effective.”
>
> — Carl Pritchard

Theoretical studies on the information structure-prosody interface have stated for some time that there is a correspondence between how the linguistic content is structured communicatively and how intonation is used in human speech to convey that content. In the present dissertation, this correspondence (in particular, the relationship between hierarchical thematicity and prosodic variation) has been brought to the foreground from an empirical perspective in the context of expressive speech generation. Corpus-based experiments and data-driven implementations support initial expectations on the potential of the information structure–prosody interface applied to speech technologies. The use of this potential is an initial step ahead in communicative approaches for prosody generation within TTS/CTS applications that is one of the key aspects for a next generation of more expressive conversational virtual agents.

This final chapter is organized as follows: in Section 7.1, the conclusions that can be drawn from the dissertation are outlined; Section 7.2 summarizes its main contribution; Section 7.3 introduces future work; finally, the list of publications during these doctoral years is introduced in Section 7.4.

7.1. Conclusions

In this dissertation, empirical evidence for the interaction between information structure and prosody has been provided. Several key aspects have been demonstrated, namely:

- speakers from different dialectal areas show commonalities in expressing thematicity by means of prosody in a corpus of read speech in English, despite the fact that a certain degree of variation (especially in F0) is also found;
- a tripartite hierarchical thematicity is more convenient for speech prosody generation than binary flat theme–rheme descriptions, especially for long complex sentences;
- mean values of normalized acoustic parameters (in particular, z-scores for F0, intensity and speech rate) show distinct distributions across hierarchical thematicity labels;
- a characterization of hierarchical thematicity using three acoustic elements allows the automatic generation of prosody enrichment using SSML tags in TTS applications;
- perception tests underline the importance of speech rate and a combination of prosodic cues (namely, F0, breaks and speech rate) to signal thematicity, which can be exploited for a more varied, and thus more expressive prosody generation of the synthesized speech.

The present dissertation is a proof of concept of the applicability of the information structure–prosody interface in speech synthesis, but there are many issues that remain unexplored. For instance, only thematicity at the sentence level has been investigated. Other dimensions of the communicative structure (like givenness and focus, as defined by Mel'čuk (2001)) may also have a strong correspondence with prosody. With respect to the experimental setup, the corpus is rather limited in size and register. Results from classification experiments proved that embedded themes (e.g., T1(SP1), T1(R1)) and rhemes (e.g., R1(SP1), R1(R1)) are confused with level 1 themes (T1) and rhemes (R1) respectively. This may be due to the lack of representativeness of the corpus, or it may indicate that themes share acoustic properties independently of their level of embeddedness. A larger corpus (with a balanced amount of classes) needs to be compiled in order to prove whether there are significant differences between embedded spans, depending on which thematicity span they are embedded into.

With respect to prosody, it has been proved in classification experiments that ToBI can be mapped to three prosodic elements. These results have been further tested in a rule-based approach for the automatic labeling of prosodic phrases

and in the generation of thematicity-based prosody enrichment in a CTS application. Despite the fact that a prosody representation based on mean normalized acoustic values does not suffice to address the requirements for prosody modeling in a pre-processing stage for TTS applications, the analysis of prosodic phrases is convenient for establishing the connection to thematicity and the derivation of acoustic values for prosody enrichment using SSML tags.

7.2. Contribution of the Dissertation

This dissertation contributes both from a theoretical and technical perspective to state-of-the-art research on the integration of the information structure–prosody interface in speech technologies. The theoretical contribution of this dissertation unfolds around the empirical findings that confirm and expand existing theories on the information structure–prosody correspondence and the way this correspondence can be transferred to speech prosody generation within a speech synthesis application. The technical contribution of this dissertation consists of two applications:

1. a tool for automatic prosody segmentation, i.e., the automatic prosody tagger, deployed on an extension of the Praat Software for feature annotation, presented as the web interface *Praat on the Web*;

2. a tool for prosody enrichment based on hierarchical thematicity within a CTS application.

The automatic prosody tagger is devised as a benchmark to test in practice empirical results from the relationship found between prosodic units and acoustic parameters. Results from manual annotation and classification experiments are used to create a set of rules that segment speech files into prosodic phrases and mark prominence within these segments. The output of the tagger has been proved to be applicable in the context of spontaneous speech prosody tagging of prosodic phrases across different languages (i.e., Spanish and English). This technical contribution fosters further research in the area of speech prosody to describe other prosodic units (e.g., prosodic words) in terms of acoustic parameters and is, furthermore, scalable to other registers and languages.

On the other hand, the tool for prosody enrichment allows testing of empirical findings on the information structure–prosody correspondence in a speech synthesis setting. The main contribution of this tool is the application of data-driven approaches for the generation of expressive prosody based upon a formal representation of thematicity. Moreover, it involves a change of perspective that mo-

tivates further research into communicatively-oriented approaches for TTS/CTS applications.

7.3. Future Work

This dissertation opens up a wide spectrum of research lines from different perspectives. As already mentioned, the first follow-up involves the analysis of more dimensions of communicative structure in connection with prosody. This includes the compilation of corpora annotated with communicative structure. Secondly, the implementation of prosody generation as a pre-process instead of a post-process is, probably, the most interesting next task to be carried out as a way to overcome inherited errors from the prosody generation module of the TTS, which could not be overridden when SSML tags were applied to the default synthesis. With respect to the development of tools for the annotation of speech prosody, an expansion of the prosody tagger is foreseen to identify and tag other (larger and smaller) prosodic units. Apart from these tasks, there are some research lines that have already been initiated in collaboration with experts in different areas that are outlined in the following sections. Section 7.3.1 sketches the work on the correspondence between information structure and multimodal analysis and generation. Section 7.3.2 outlines the development towards the annotation of thematicity in spontaneous speech and in other languages than English.

7.3.1. Information Structure in Multimodal Analysis and Generation

Having proved that there is a correspondence between information structure and prosody on the one hand, and between prosody and gestures on the other hand (see e.g., (Prieto et al., 2015)), it is reasonable to think that there could be also a correspondence between information structure and non-verbal behavior (i.e., facial expressions and gestures). The integration of verbal and non-verbal communication can be beneficial in both analysis and generation scenarios. In cooperation with the CM-TECH lab from the DTIC-UPF, the relationship between thematicity and facial expressions is currently under investigation.

Within the framework of the KRISTINA project[1], I am participating in the annotation of multimodal non-verbal cues in the five working languages of the project (Spanish, German, Polish, Turkish and Arabic). A set of guidelines was established for the annotation of non-verbal behavior. These guidelines involve

[1] http://kristina-project.eu/en/

the definition a discrete scale of values within the valence–arousal space to represent emotional states conveyed by the combination of facial expressions, gestures and voice. The corpus used for annotation consists in spontaneous dialogs around relevant topics for the use cases in the project (Wanner et al., 2016, 2017). A further line of improvement in this direction is to integrate results from this analysis of non-verbal behavior to the verbal content and explore whether there is a correspondence between information structure and facial expressions and whether this relationship varies across languages and cultures.

7.3.2. Annotation of Thematicity in Spontaneous Speech

The corpus that was recorded for this dissertation also contains a small spontaneous speech sample by each participant. These samples are being annotated with thematicity and guidelines for the annotation of spontaneous speech are being developed. The annotation of spontaneous speech differs greatly from the annotation of written texts, and involves taking decisions around linguistic events that are inherent to spontaneous speech, such as hesitations, filled pauses, truncation, reformulation, etc. For instance, one of the issues we are facing in this context is coming to an agreement upon what a main proposition is. This is not a problem in texts, as fullstops clearly indicate the end of a sentence, and consequently, the end of a main proposition. In spontaneous speech, however, establishing when a sentence ends is not as straight forward as it may seem. Thematicity annotation over whole propositions is also being carefully looked into. The function of specifiers in spontaneous speech is investigated from a closer perspective as well. In this respect, different types of specifiers are being identified; e.g., those including copulative conjunctions, such as “and”, which may differ in function depending on the way they are uttered in spontaneous monologues.

This task is carried out in collaboration with Dr. Alicia Burga and Beatriz Fisas under the supervision of Dr. Leo Wanner and precious pieces of advice are eventually given by Professor Igor Mel’čuk. The annotation of thematicity for spontaneous speech is foreseen to be expanded to different genres (e.g., dialogs and story telling) and languages apart from English (initially, Spanish and German). Even though it is a laborious (and rather slow) task, the availability of annotated resources including thematicity is one of the key points for further development and integration of the information structure–prosody interface applied to speech technologies.

7.4. List of Publications

- Domínguez, M., Farrús, M., Burga, A., and Wanner, L. (2014a). The Information Structure–Prosody Language Interface Revisited. In *Proceedings of the 7th International Conference on Speech Prosody*. Dublin, Ireland, pages 539–543.

- Domínguez, M., Farrús, M., Burga, A., and Wanner, L. (2014b). Towards automatic extraction of prosodic patterns for speech synthesis. In *Proceedings of the 7th International Conference on Speech Prosody*. Dublin, Ireland, pages 1105–1109.

- Domínguez, M., Farrús, M., Burga, A., and Wanner, L. (2016a). Using hierarchical information structure for prosody prediction in content-to-speech applications. In *Proceedings of the 8th International Conference on Speech Prosody*. Boston, USA, pages 1019–1023.

- Domínguez, M., Farrús, M., and Wanner, L. (2016b). Combining acoustic and linguistic features in phrase-oriented prosody prediction. In *Proceedings of the 8th International Conference on Speech Prosody*. Boston, USA, pages 796–800.

- Domínguez, M., Farrús, M., and Wanner, L. (2016c). An automatic prosody tagger for spontaneous speech. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*. Osaka, Japan, pages 377–387.

- Domínguez, M., Latorre, I., Farrús, M., Codina, J., and Wanner, L. (2016d). Praat on the Web: An upgrade of Praat for semiautomatic speech annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. Osaka, Japan, pages 218–222.

- Sukno, F.M., Domínguez, M., Ruiz, A., Schiller, D., Lingenfelser, F., Pragst, L., Kamateri, E., Vrochidis, S. (2016e). A multimodal annotation schema for non-verbal affective analysis in the health-care domain. In *Proceedings of the 1st International Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction (MARMI 2016)*. New York, USA. New York: ACM, 2016, pages 9-14.

- Wanner, L., Blat, J., Dasiopoulou, S., Domínguez, M., Llorach, G., Mille, S., Sukno, F., Kamateri, E., Kompatsiaris, I., Vrochidis, S., André, E., Lingenfelser, F., Mehlmann, G., Stam, A., Stellingwerff, L. (2016f). Towards

a multimedia knowledge-based agent with social competence and human interaction capabilities. In *Proceedings of the 1st International Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction (MARMI 2016)*. New York, USA. New York: ACM, pages 21-26.

- Domínguez, M., Farrús, M., and Wanner, L. (2017). A thematicity-based prosody enrichment tool for CTS. In *INTERSPEECH2017: Show and tell demonstrations*, Stockholm, Sweden.

- Domínguez, M., Burga, A, Farrús, M., and Wanner, L. Compilation of corpora to study the information structure–prosody interface. Submitted to *11th edition of the Language Resources and Evaluation Conference (LREC2018)*, Miyazaki, Japan.

- Domínguez, M., Farrús, M., and Wanner, L. A data-driven approach to thematicity-based prosody enrichment. For submission to *Speech Prosody (SP2018)*, Poznań, Poland.

- Domínguez, M., Burga, A., Farrús, M., and Wanner, L. The information structure–prosody interface: an empirical approach for concept-to-speech applications. In preparation for submission to *Computer Speech and Language Journal*.

Bibliography

Ammann, H. (1928). *Die Menschliche Rede. Der Satz*. Darmstadt: Lahr.

Anumanchipalli, G. (2013). *Intra-Lingual and Cross-Lingual Prosody Modelling*. PhD thesis, Carnegie Mellon University.

Audibert, N., Aubergé, V., and Rilliard, A. (2005). The relative weights of the different prosodic dimensions in expressive speech: A resynthesis study. In Tao, J., Tan, T., and Picard, R. W., editors, *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction, ACII 2005*, pages 527–534, Beijing, China. Springer Berlin Heidelberg.

Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2008). ANALOR. a tool for semi-automatic annotation of french prosodic structure. In *Proceedings of the International Conference on Speech Prosody*, pages 119–122.

Ballesteros, M., Bohnet, B., Mille, S., and Wanner, L. (2015). Data-driven sentence generation with non-isomorphic trees. In *Proceedings of the Annual Conference of the North American Association for Computational Linguistics – Human Language Technologies (NAACL – HLT)*.

Batliner, A. and Möbius, B. (2005). Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground? In *The integration of phonetic knowledge in speech technology*, volume 25, pages 21–44, Netherlands. Springer.

Baumann, S. (2012). *The Intonation of Givenness. Evidence from German*. Max Niemeyer Verlag, Berlin, Boston.

Baumann, S., Brinckmann, C., Hansen-Schirra, S., Kruijff, G., Kruijff-Korbayová, I., Neumann, S., and Teich, E. (2004). Multi-dimensional annotation of linguistic corpora for investigating information structure. *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 39–46.

Beckman, M. E., Hirschberg, J. B., and Shattuck-Hufnagel, S. (2004). The Original ToBI System and the Evolution of the ToBI Framework. In Jun, S., editor, *Prosodic Models and Transcription: Towards Prosodic Typology*, pages 9–54. Oxford University Press.

Beckman, M. E. and Pierrehumbert, J. (1986). Intonational Structure in Japanese and English. *Phonology Yearbook*, 3:255–310.

Bierner, G. (1998). TraumaTalk: content-to-speech generation for decision support at point of care. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 698–702.

Black, A. W. and Taylor, P. A. (1997). The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK. Avaliable at http://www.cstr.ed.ac.uk/projects/festival.html.

Bock, J. K., Mazzella, J. R., and Bock, K. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, 11(1):64–76.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [Computer program], http://www.praat.org/, version 6.0.14.

Bohnet, B., Burga, A., and Wanner, L. (2013). Towards the annotation of penn treebank with information structure. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1250–1256, Nagoya, Japan.

Bouayad-Agha, N., Casamayor, G., Mille, S., and Wanner, L. (2012). Perspective-Oriented Generation of Football Match Summaries: Old Tasks, New Challenges. *ACM Transactions on Speech and Language Processing*, 9(2).

Brown, G. (1983). Prosodic structure and the given/new distinction. In Cutler, A. and Ladd, D. R., editors, *Prosody: Models and Measurements*, pages 67–77. Springer, Berlin, Heidelberg.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics & Philosophy*, 26(5):511–545.

Büring, D. (2016). *Intonation and Meaning*. Oxford University Press, Oxford.

Calhoun, S. (2010). The centrality of metrical structure in signalling information structure: A probabilistic perspective. *Language*, 1(86):1–42.

Campbell, N. and Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2417–2420, Barcelona, Spain.

Cannam, C., Landone, C., and Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy.

Chafe, W. L. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, Chicago and London.

Chafe, W. L. and Li, C. N. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view in subject and topic. *Subject and Topic*, pages 25–55.

Charniak, E. and al., E. (2000). *BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43.* Linguistic Data Consortium, Philadelphia.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Christodoulides, G. (2014). Praaline: Integrating tools for speech corpus research. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Clark, H. H. and Haviland, S. E. (1977). Comprehension and the given-new contract. *Discourse production and comprehension. Discourse processes: Advances in research and theory*, 1:1–40.

Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, pages 37–46. Sage Publications Inc.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, v., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6).*

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2).

Daneš, F. (1970). One instance of Prague School methodology: Functional analysis of utterance and text. *Garvin*, pages 132–141.

de Carolis, B., Pelachaud, C., Poggi, I., and Steedman, M. (2004). APML, a mark-up language for believable behavior generation. In Prendinger, H. and Ishizuka, M., editors, *Life-like Characters. Tools, Affective Functions and Applications*, pages 65–85. Springer.

Delmonte, R. and Tripodi, R. (2015). Semantics and discourse processing for expressive tts. In *Proceedings of the EMNLP Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, pages 32–43.

Elvira-García, W., Roseano, P., Fernández-Planas, A. M., and Martínez-Celdrán, E. (2016). A tool for automatic transcription of intonation: Eti_tobi a tobi transcriber for spanish and catalan. *Language Resources and Evaluation*, 50(4):767–792.

Erteschik-Shir, N. (2007). *Information Structure: The Syntax-Discourse Interface*. Oxford University Press, Oxford.

Féry, C. (2013). Focus as prosodic alignment. *Natural Language & Linguistic Theory*, 31(3):683–734.

Féry, C. and Ishihara, S., editors (2016). *The Oxford Handbook of Information Structure*. Oxford University Press, Oxford, UK.

Féry, C. and Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, 36(4):680–703.

Firbas, J. (1964). On defining the theme in functional sentence perspective. In *Travaux linguistiques de Prague*, pages 1267–1280.

Fowler, C. A. and Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5):489–504.

Fusijaki, H. (2012). *Prosody, Models and Spontaneous Speech*, chapter 3, pages 27–42. Springer Science & Business Media.

Grabe, E., Nolan, F., and Farrar, K. (1998). Ivie - a comparative transcription system for intonational variation in english. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 98)*, pages 1259–1262, Sydney, Australia.

Grice, H. (1989). *Studies in the way of words*. Harvard University Press.

Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. Foris, Dordrecht.

Haji-Abdolhosseini, M. and Müller, S. (2003). Constraint-Based Approach to Information Structure and Prosody Correspondence. In *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, pages 143–162. CSLI Publications.

Hajicová, E. (1986). Focussing- A Meeting Point Linguistics and Artificial Intelligence. In *Artificial Intelligence II: Methodology, Systems, Applications - Proceedings of the Second International Conference on Artificial Intelligence: Methodology, Systems, Applications, AIMSA 1986, Varna, Bulgaria, September 16-19, 1986*, pages 311–321.

Hajičova, E., Partee, B., and Sgall, P. (1998). *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Halliday, M. (1967). Notes on Transitivity and Theme in English, Parts 1-3. *Journal of Linguistics*, 3(1):37–81.

Hirose, K., Fujisaki, H., and Yamaguchi, M. (1984). Synthesis by rule of voice fundamental frequency contours of spoken japanese from linguistic information. In *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP 84)*, pages 597–600, San Diego, California, USA.

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1–2):31–43.

Hirschberg, J. (2008). *Pragmatics and Intonation*, pages 515–537. Blackwell Publishing Ltd.

Hirst, D. (2001). *Automatic Analysis of Prosody for Multi-lingual Speech Corpora*, pages 320–327. Wiley, London.

Hirst, D. (2009). The rhythm of text and the rhythm of utterances: From metrics to models. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1519–1522.

Hirst, D. and Auran, C. (2005). Analysis by synthesis of speech prosody : the prozed environment. In *Proceesings of Interspeech*, pages 3225–3228, lisboa, Portugal.

Hirst, D. and Cristo, A. d. (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press.

Holmes, J. N. (1987). From text to speech: The MITalk system.

Hualde, J. (2000). *Intonation in Spanish and the other Ibero-Romance languages: Overview and status quaestionis.*, pages 101–115. Amsterdam: Benjamins, Gainesville, Florida.

Hualde, J. and Prieto, P. (2016). Towards an international prosodic alphabet (ipra). *Journal of the Association for Laboratory Phonology*, 7(1):1–25.

Kalbertodt, J., Primus, B., and Schumacher, P. B. (2015). Punctuation, prosody, and discourse: Afterthought vs. right dislocation. *Frontiers in Psychology*, 6:1803.

Kruijff-Korbayová, I., Ericsson, S., Rodríguez, K. J., and Karagrjosova, E. (2003). Producing Contextually Appropriate Intonation in an Information-State Based Dialogue System. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 227–234.

Kügler, F., Smolibocki, B., and Stede, M. (2012). Evaluation of information structure in speech synthesis : The case of product recommender systems perception. In *ITG Conference on Speech Communication, IEEE*, pages 26–29.

Kügler, F., Smolibocki, B., Stede, M., and Varges, S. (2013). Information Structure in Speech Synthesis : Early Focus and Post-Focal Givenness. In Wagner, P., editor, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pages 56–63. TUDpress, Dresden.

Ladd, D. (1996). *Intonational Phonology*. Cambridge Studies in Linguistics. Cambridge University Press.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge University Press, Cambridge.

Lee, J. S., Kim, B., and Lee, G. G. (2002). Automatic corpus-based tone and break-index prediction using k-tobi representation. *Transactions on Asian Language Information Processing (TALIP)*, 1(3):207–224.

Li, Y., Campbell, N., and Tao, J. (2015). Voice quality: Not only about you; but also about your interlocutor. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4739–4743. IEEE.

Lindstrom, A., Bretan, I., and Ljungqvist, M. (1996). Prosody generation in text-to-speech conversion using dependency graphs. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1341–1344.

Mathesius, V. (1929). Zur Satzperspektive im modernen Englisch. In *Archiv für das Studium der neueren Sprachen und Literaturen*, volume 155, pages 202–210. Erich Schmidt Verlag.

MATLAB (2007). MATLAB Optimization Toolbox version 7.10.0.

Mel'čuk, I. (1981). Meaning-Text Models: A Recent Trend in Soviet Linguistics. *Annual Review of Anthropology*, 10:27–62.

Mel'čuk, I. A. (2001). *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Benjamins, Amsterdam, Philadephia.

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In *Proceedings of the 2nd International Conference on Speech Prosody*, pages 549–552, Nara, Japan.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, TIWTE '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mixdorff, H. (2002). Speech technology, tobi, and making sense of prosody. In *Proceedings of the First International Conference on Speech Prosody*, pages 31–37, Aix-en-Provence, France.

Mixdorff, H. (2015). *Extraction, Analysis and Synthesis of Fujisaki model Parameters*, pages 35–47. Springer, Berlin, Heidelberg.

Nespor, I. and Vogel, I. (1986). *Prosodic Phonology*. Foris, Dordrecht.

Noguchi, H., Kiriyama, K., Matsuda, H., Taniguchi, M., Den, Y., and Katagiri, Y. (1999). Automatic labeling of japanese prosody using j-tobi style description. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999*, Budapest, Hungary.

Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In *The Handbook of Phonetic Sciences*. Blackwell Publishers Ltd, Oxford.

Olaszy, G. and Nemeth, G. (1997). Prosody generation for German CTS/TTS systems (from theoretical intonation patterns to practical realisation). *Speech Communication*, 21:37–60.

Osgood, C. (1960). What is a language? In Aaronson, D. and Rieber, R., editors, *Psycholinguistic Research: Implications and Applications*, pages 189–228. Erlbaum, Hillsdale, New Jersey.

Pfaff, B. (2015). *PSPP Users' Guide. GNU PSPP Statistical Analysis Software. Release 0.8.5.* Free Software Foundation, Boston, MA.

Prevost, S. (1996). An information structural approach to spoken language generation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 294–301, Morristown, NJ, USA. Association for Computational Linguistics.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90(6):2956–2970.

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., and Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49:41–54.

Prom-On, S., Xu, Y., Gu, W., Arvaniti, A., Nam, H., and Whalen, D. H. (2016). The common prosody platform (cpp): Where theories of prosody can be directly compared. In *Proceedings of the 8th International Conference on Speech Prosody*, pages 1–5, Boston, USA.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Romanelli, M., Baumann, S., Kruijff-Korbayova, I., and Kruijff, G.-J. (2001). Modeling givenness and contrast in MARY (Modular Architecture for Research on speech sYnthesis). Technical report, Ms. Saarland University.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Rose, P. (2002). *Forensic Speaker Identification*. Taylor & Francis, New York, London.

Rosenberg, A. (2010). AutoBI - A tool for automatic ToBI annotation. In *Proceedings of Interspeech*, pages 146–149, Makuhari, Japan.

Salvo Rossi, P., Palmieri, F., and Cutugno, F. (2002). A method for automatic extraction of fujisaki-model parameters. In *Speech Prosody*, pages 615–618, Aix-en-Provence, France.

Schröder, M. and Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6(4):365–377.

Schwarzschild, R. (1999). Givenness, avoidf and other constraints on the placement of accent. *Natural Language Semantics*, 7(1):141–177.

Schweitzer, A., Braunschweiler, N., Dogil, G., Klankert, T., Möbius, B., Möhler, G., Morais, E., Säuberlich, B., and Thomae, M. (2006). *Multimodal Speech Synthesis*. Springer.

Selkirk, E. O. (1984). *Phonology and Syntax: The relation between sound and structure*. The MIT Press, Cambridge, Massachussetts.

Sgall, P. (2000). Functional sentence perspective in written and spoken communication. Studies in English Language. *Journal of Pragmatics*, 32(5):639–644.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Interspeech 2005, 9th European Conference on Speech Communication and Technology*, pages 1781–1784, Lisbon, Portugal.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. In *2nd International Conference on Spoken Language Processing (ICSLP 92)*, pages 867–870, Banff, Canada.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (2010). ToBI: A standard for labeling English prosody. In *Proceedings of Interspeech*, pages 146–149, Makuhari, Japan.

Stede, M. and Mamprin, S. (2016). Information structure in the Potsdam Commentary Corpus: Topics. In *Tenth International Conference on Language Resources and Evaluation, LREC*, pages 1718–1723.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic inquiry*, 31(4):649–689.

Steedman, M. (2004). Using APML to specify intonation. *Unpublished tutorial paper*, pages 1–24.

Steedman, M. (2013). The surface-compositional semantics of english intonation. *Language*, 90:2–57.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: A web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Styler, W. (2013). *Using Praat for Linguistic Research*. University of Colorado at Boulder Phonetics Lab, Colorado, USA.

Tabet, Y. and Boughazi, M. (2011). Speech synthesis techniques. A survey. In *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, pages 67–70.

Tahon, M. and Devillers, L. (2016). Towards a small set of robust acoustic features for emotion recognition: Challenges. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):16–28.

Taylor, P. (1998). The tilt intonation model. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, pages 1383–1386.

Taylor, P. and Isard, A. (1997). Ssml: A speech synthesis markup language. *Speech Communication*, 21(1-2):123–133.

Tsai, C.-Y., Kuo, C.-K., Wang, Y.-R., Chen, S.-H., Liao, I.-B., and Chiang, C.-Y. (2014). Hierarchical prosody modeling of English speech and its application to TTS. In *17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–6. IEEE.

Tsatsaronis, G., Varlamis, I., and Nørvåg, K. (2012). Semafor: Semantic document indexing using semantic forests. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1692–1696, New York, NY, USA. ACM.

Tseng, C. (2004). Intensity in relation to prosody organization. In *International Symposium on Chinese Spoken Language Processing*, pages 217–220. IEEE.

van Donselaar, W. and Lentz, J. (1994). The function of sentence accents and given/new information in speech processing: different strategies for normal-hearing and hearing-impaired listeners? *Language and Speech*, 37(4):375–391.

Van Santen, J. P., Sproat, R., Olive, J., and Hirschberg, J. (2013). *Progress in speech synthesis*. Springer Science & Business Media.

Vanrell, M., Mascaró, I., Torres-Tamarit, F., and Prieto, P. (2013). Intonation as an Encoder of Speaker Certainty: Information and Confirmation Yes-No Questions in Catalan. *Language and Speech*, 56(2):163–190.

von Heusinger, K. (1999). *Intonation and Information Structure. The Representation of Focus in Phonology*. PhD thesis, University of Kostanz.

Von Stechow, A. (1981). *Topic, Focus and Local Relevance*, pages 95–130. Springer Netherlands, Dordrecht.

Wanner, L., André, E., Blat, J., Dasiopoulou, S., Farrús, M., Fraga, T., Kamateri, E., Lingenfelser, F., Llorach, G., Martínez, O., Meditskos, G., Mille, S., Minker, W., Pragst, L., Schiller, D., Stam, A., Stellingwerff, L., Sukno, F., Vieru, B., and Vrochidis, S. (2017). KRISTINA: A Knowledge-Based Virtual Conversation Agent. In *Proceedings of the 15th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS)*, Oporto, Portugal.

Wanner, L., Blat, J., Dasiopoulou, S., Domínguez, M., Llorach, G., Mille, S., Sukno, F., Kamateri, E., Vrochidis, S., Kompatsiaris, I., et al. (2016). Towards a multimedia knowledge-based agent with social competence and human interaction capabilities. In *Proceedings of the 1st International Workshop on Multimedia Analysis and Retrieval for Multimodal Interaction*, pages 21–26. ACM Digital Library.

Wanner, L., Bohnet, B., and Giereth, M. (2003). Deriving the Communicative Structure in Applied NLG. In *Proceedings of the 9th European Workshop on Natural Language Generation at the Biannual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 100–104.

Watts, O., Henter, G. E., Merritt, T., Wu, Z., and King, S. (2016). From HMMS to DNNS: Where do the improvements come from? In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5505–5509. IEEE.

Wennerstrom, A. (2001). *The Music of Everyday Speech. Prosody and Discourse Analysis*. Oxford University Press, Oxford.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 27(1):55–105.

Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences*, 1(1):85–115.

Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, pages 7–10, Aix-en-Provence, France.

Xydas, G., Spiliotopoulos, D., and Kouroupetroglou, G. (2005). Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora. *IEICE transactions on information and systems*, 88(3):510–518.

Zubizarreta, M. L. (2016). *Nuclear Stress and Information Structure*, pages 1–40. Oxford University Press.

Appendices

Appendix A

CORPUS: RAW TEXT

1. Ms. Haag plays Elianti.
2. Rolls-Royce Motor Cars Inc. said it expects its U.S. sales to remain steady at about 1,200 cars in 1990.
3. The luxury auto maker last year sold 1,214 cars in the U.S.
4. BELL INDUSTRIES Inc. increased its quarterly to 10 cents from seven cents a share.
5. The new rate will be payable Feb. 15.
6. A record date hasn't been set.
7. Bell, based in Los Angeles, makes and distributes electronic, computer and building products.
8. The proposed changes also would allow executives to report exercises of options later and less often.
9. "Apparently the commission did not really believe in this ideal."
10. But about 25% of the insiders, according to SEC figures, file their reports late.
11. The SEC will probably vote on the proposal early next year, he said.
12. Not all those who wrote oppose the changes.
13. According to some estimates, the rule changes would cut insider filings by more than a third.

14. The SEC's Mr. Lane vehemently disputed those estimates.

15. The proposed rules also would be tougher on the insiders still required to file reports, he said.

16. Companies would be compelled to publish in annual proxy statements the names of insiders who fail to file reports on time.

17. Many investors wrote asking the SEC to require insiders to report their purchases and sales immediately, not a month later.

18. Investors who want to change the required timing should write their representatives in Congress, he added.

19. Both funds are expected to begin operation around March 1, subject to Securities and Exchange Commission approval.

20. For its employees to sign up for the options, a college also must approve the plan.

21. Some 4,300 institutions are part of the pension fund.

22. The new "social choice" fund will shun securities of companies linked to South Africa, nuclear power and in some cases, Northern Ireland.

23. Also excluded will be investments in companies with "significant" business stemming from weapons manufacture, alcoholic beverages or tobacco.

24. Sixty percent of the fund will be invested in stocks, with the rest going into bonds or short-term investments.

25. The bond fund will invest in high-grade or medium-grade bonds, mortgages or asset-backed securities, including as much as 15% in foreign securities.

26. The fund also might buy and sell futures and options contracts, subject to approval by the New York State Insurance Department.

27. The investment choices offered by the pension fund currently are limited to a stock fund, an annuity and a money-market fund.

28. The company said the plan, under review for some time, will protect shareholders against abusive takeover tactics.

29. W. Ed Tyler, 37 years old, a senior vice president at this printing concern, was elected president of its technology group, a new position.

30. The oboist Heinz Holliger has taken a hard line about the problem:

31. Richard Stoltzman has taken a gentler, more audience-friendly approach.

32. But you can't dismiss Mr. Stoltzman's music or his motives as merely commercial and lightweight.

33. He believes in what he plays, and he plays superbly.

34. His recent appearance at the Metropolitan Museum, dubbed "A Musical Odyssey", was a case in point.

35. It felt more like a party, or a highly polished jam session with a few friends, than a classical concert.

36. He launched into Saint-Saens's "The Swan" from "Carnival of the Animals," a favorite encore piece for cellists, with lovely, glossy tone and no bite.

37. Mr. Stoltzman introduced his colleagues: Bill Douglas, pianist/bassoonist/-composer and an old buddy from Yale, and jazz bassist Eddie Gomez.

38. Bach's "Air" followed.

39. "Deep Peace" also featured a slide show of lovely but predictable images of clouds, beaches, deserts, sunsets, etc.

40. That went over the permissible line for warm and fuzzy feelings.

41. Was this why some of the audience departed before or during the second half?

42. Or was it because Ms. Collins had gone?

43. Mr. Reich's new "Different Trains" for string quartet uses the technique magisterially.

44. Mr. Stoltzman must have worried that his audience might not be able to take it: He warned us in advance that "New York Counterpoint" lasts $11\frac{1}{2}$ minutes.

45. Is this the future of chamber music?

46. What's next?

47. Slides to illustrate Shostakovich quartets?

48. But it was neither deep nor lasting: light entertainment that was no substitute for an evening of Brahms.

49. Ms. Waleson is a free-lance writer based in New York.

50. In fact, he liberated the U.S. from one of the world's most corrupt organizations – UNESCO.

51. This is the U.N. group that managed to traduce its own charter of promoting education, science and culture.

52. Ever since, the remaining members have been desperate for the United States to rejoin this dreadful group.

53. Now UNESCO apologists are lobbying President Bush to renege on President Reagan's decision to depart.

54. The Orwellian "New World Information Order" would give government officials rights against the press.

55. UNESCO somehow converted the founding U.N. ideals of individual rights and liberty into "peoples' rights."

56. UNESCO is now holding its biennial meetings in Paris to devise its next projects.

57. Mr. Mayor's hope that references to "press freedom" would survive unamended seems doomed to failure.

58. The current phrasing is "educating the public and media to avoid manipulation."

59. He hasn't been able to replace the M'Bow cabal.

60. Other countries, including West Germany, may have a hard time justifying continued membership.

61. The Babelists of the United Nations are experts at obfuscation.

62. The researchers said they have isolated a plant gene that prevents the production of pollen.

63. On a commercial scale, the sterilization of the pollen-producing male part has only been achieved in corn and sorghum feed grains.

64. In a labor-intensive process, the seed companies cut off the tassels of each plant, making it male sterile.

65. They sow a row of male-fertile plants nearby, which then pollinate the male-sterile plants.

66. The vast majority of the U.S. corn crop now is grown from hybrid seeds produced by seed companies.

67. A similar technique is almost impossible to apply to other crops, such as cotton, soybeans and rice.

68. Mr. Leemans said this genetic manipulation doesn't hurt the growth of that plant.

69. They attached a second gene, for herbicide resistance, to the pollen-inhibiting gene.

70. Both genes are then inserted into a few greenhouse plants, which are then pollinated and allowed to mature and produce seed.

71. One technique developed by some of these companies involves a chemical spray supposed to kill only a plant's pollen.

72. "There is a large market out there hungry for hybrid seeds", he said.

73. Nevertheless, he said, he is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton.

74. Already, the consequences are being felt by other players in the financial markets – even governments.

75. What triggered the latest clash was a skirmish over the timing of a New Zealand government bond issue.

76. The dispute shows clearly the global power of Japan's financial titans.

77. Aside from Nomura's injured pride, the biggest victim so far has been the New Zealand government.

78. New Zealand's finance minister, David Caygill, lashed out at such suggestions.

79. "It may very well be what the Japanese banks want," he told Radio New Zealand.

80. Both sides are jealously guarding their turf, and relations have been at a flashpoint for months.

81. The banks badly want to break into all aspects of the securities business.

82. And their suspicions of each other run deep.

83. In the past year, both have tried to stretch the limits of their businesses.

84. Mr. Conlon was executive vice president and director of the equity division of the international division of Nikko Securities Co.

85. As Yogi Berra might say, it's deja vu all over again.

86. "Old-time kiddies," he says.

87. But for the next few months, these boys of summers long past are going to be reveling in an Indian summer of the soul.

88. Now that the baseball season is officially over, you see, it's time for a new season to begin.

89. "Someone always makes you quit," says legendary St. Louis Cardinals centerfielder Curt Flood, the league's commissioner.

90. "You feel you want one more –one more at-bat, one more hit, one more game."

91. When he sent letters offering 1,250 retired major leaguers the chance of another season, 730 responded.

92. For some players, the lure is money –up to $15,000 a month.

93. Others, just released from the majors, hope the senior league will be their bridge back into the big-time.

94. (No one has worked out the players' average age, but most appear to be in their late 30s).

95. "There will be a lot of malice."

96. Men who have played hard all their lives aren't about to change their habits, he says.

97. "If you know how to slide, it's no problem," he says.

98. After all, he says, "Even to make love, you need experience.

99. Stewart & Stevenson Services Inc. said it received two contracts totaling $19 million to build gas-turbine generators.

100. Statistics Canada said service-industry output in August rose 0.4% from July.

101. Both General Motors Corp. and Ford Motor Co. have been trying to amass 15% stakes in Jaguar.

102. Many investors certainly believe a bidding war is imminent.

103. Such a countermove could end Jaguar's hopes for remaining independent and British-owned.

104. Dow will own 60% of the venture, with Eli Lilly holding the rest.

105. The 45-year-old Mr.Kuehn, who has a background in crisis management, succeeds Alan D. Rubendall, 45.

106. Mr.Kuehn, the company said, will retain the rest of the current management team.

107. The gains also sparked buying interest in other real-estate companies, traders said.

108. The balance of short positions outstanding fell 159.7 billion yen, to 779.8 billion yen.

109. No one wants stock on their books.

Appendix B

CORPUS: TEXT ANNOTATED WITH THEMATICITY

1. [Ms. Haag]T1 [plays Elianti]R1.
2. [[Rolls - Royce Motor Cars Inc.]T1(SP1) [said]R1(SP1)]SP1 [it]T1 [expects its U.S. sales to remain steady at about 1,200 cars in 1990]R1.
3. [The luxury auto maker]T1 [last year sold 1,214 cars in the U.S]R1.
4. [BELL INDUSTRIES Inc.]T1 [increased its quarterly to 10 cents from seven cents a share]R1.
5. [The new rate]T1 [will be payable Feb. 15]R1.
6. [A record date]T1 [has n't been set]R1.
7. [Bell , based in Los Angeles]T1 , [makes and distributes electronic , computer and building products]R1.
8. [The proposed changes]T1 [also would allow executives to report exercises of options later and less often]R1.
9. " [Apparently]SP1 [the commission]T1 [did not really believe in this ideal]R1. "
10. [But]SP1 [about 25% of the insiders]T1 , [according to SEC figures]SP2 , [file their reports late]R1.
11. [The SEC]T1 [will probably vote on the proposal early next year]R1 , [[he]T1(SP1) [said]R1(SP1)]SP1.
12. [Not all those {[who]T1(P2) [wrote]R1(P2)}P2]T1 [oppose the changes]R1.

13. [According to some estimates]SP1 , [the rule changes]T1 [would cut insider filings by more than a third]R1.

14. [The SEC 's Mr. Lane]T1 [vehemently disputed those estimates]R1.

15. [The proposed rules]T1 [also would be tougher on the insiders still required to file reports]R1 , [[he]T1(SP1) [said]R1(SP1)]SP1.

16. [Companies]T1(P1) [would be compelled to publish in annual proxy statements the names of insiders {[who]T1(P2) [fail to file reports on time]R1(P2)}P2]R1(P1).

17. [Many investors wrote]T1 [[asking the SEC to require insiders]T1(R1) [to report their purchases and sales immediately , not a month later]R1(R1)]R1.

18. [Investors {[who]T1(P2) [want to change the required timing]R1(P2)}P2]T1 [should write their representatives in Congress]R1 , [[he]T1(SP1) [added]R1(SP1)]SP1.

19. [Both funds]T1 [are expected to begin operation around March 1 , subject to Securities and Exchange Commission approval]R1.

20. [[For its employees]T1(T1) [to sign up for the options]R1(T1)]T1 , [[a college]T1(R1) [also must approve the plan]R1(R1)]R1.

21. [Some 4,300 institutions]R1 [are part of the pension fund]T1.

22. [The new " social choice " fund]T1 [will shun securities of companies linked to South Africa , nuclear power and in some cases , Northern Ireland]R1.

23. [[Also excluded]Foc will be investments in companies with " significant " business stemming from weapons manufacture , alcoholic beverages or tobacco]R1.

24. [Sixty percent of the fund]T1 [will be invested in stocks , with the rest going into bonds or short - term investments]R1.

25. [The bond fund]T1 [will invest in high - grade or medium - grade bonds , mortgages or asset - backed securities , including as much as 15% in foreign securities]R1.

26. [The fund]T1 [also might buy and sell futures and options contracts , subject to approval by the New York State Insurance Department]R1.

27. [The investment choices offered by the pension fund]T1 [currently are limited to a stock fund , an annuity and a money - market fund]R1.

28. [[The company]T1(SP1) [said]R1(SP1)]SP1 [the plan , under review for some time]T1 , [will protect shareholders against abusive takeover tactics]R1.

29. [W. Ed Tyler , 37 years old , a senior vice president at this printing concern]T1 , [was elected president of its technology group , a new position]R1.

30. [The oboist Heinz Holliger]T1 [has taken a hard line about the problem]R1:

31. [Richard Stoltzman]T1 [has taken a gentler , more audience - friendly approach]R1

32. [But]SP1 [you ca n't dismiss]R1-1 [Mr. Stoltzman 's music or his motives]T1 [as merely commercial and lightweight]R1-2.

33. {[He]T1(P2) [believes in {[what]T1(P3) [he plays]R1(P3)}P3]R1(P2)}P2 , {[and]SP1(P3) [he]T1(P3) [plays superbly]R1(P3)}P3.

34. [His recent appearance at the Metropolitan Museum , dubbed " A Musical Odyssey]T1 , " [was a case in point]R1.

35. [It]T1 [felt more like a party , or a highly polished jam session with a few friends , than a classical concert]R1.

36. [He]T1 [launched into Saint - Saens 's " The Swan " from " Carnival of the Animals, " a favorite encore piece for cellists , with lovely , glossy tone and no bite]R1.

37. [Mr. Stoltzman]T1 [introduced his colleagues : Bill Douglas , pianist / bassoonist / composer and an old buddy from Yale , and jazz bassist Eddie Gomez]R1.

38. [Bach 's " Air "]T1 [followed]R1.

39. [" Deep Peace "]T1 [also featured a slide show of lovely but predictable images of clouds , beaches , deserts , sunsets , etc]R1.

40. [That]T1 [went over the permissible line for warm and fuzzy feelings]R1.

41. [Was this]T1 [why {[some of the audience]T1(P2) [departed before or during the second half]R1(P2)}P2]R1 ?

42. [Or]SP1 [was it]T1 [because {[Ms. Collins]T1(P2) [had gone]R1(P2)}P2]R1 ?

43. [[Mr. Reich 's new " Different Trains " for string quartet]T1 [uses the technique magisterially]R1.]Backgr

44. [Mr. Stoltzman]T1 [must have worried that {[his audience]T1(P2) [might not be able to take it]R1(P2)}P2]R1.

45. [He]T1 [warned us in advance that {[“ New York Counterpoint ”]T1(P2) [lasts 11 1/2 minutes]R1(P2)}P2]R1.

46. [What ’s]R1 [next ?]T1

47. [Slides to illustrate Shostakovich quartets ?]R1

48. [But]SP1 [it]T1 [was neither deep nor lasting]R1 : [light entertainment {[that]T1(P2) [was no substitute for an evening of Brahms]R1(P2)}P2]R2.

49. [Ms. Waleson]T1 [is a free - lance writer based in New York]R1.

50. [In fact]SP1 , [he]T1 [liberated the U.S. from [one of the world ’s most corrupt organizations]A1 – [UNESCO]A2]R1.

51. [This]T1 [is the U.N. group {[that]T1(P2) [managed to traduce its own charter of promoting education , science and culture]R1(P2)}P2]R1.

52. [Ever since]SP1 , [the remaining members]T1 [have been desperate for {[the United States]T1(P2) [to rejoin this dreadful group]R1(P2)}P2]R1.

53. [Now]SP1 [UNESCO apologists]T1 [are lobbying President Bush to renege on President Reagan ’s decision to depart]R1.

54. [The Orwellian “ New World Information Order ”]T1 [would give government officials rights against the press]R1.

55. [UNESCO]T1 [somehow converted the founding U.N. ideals of individual rights and liberty into “ peoples ’ rights]R1. ”

56. [UNESCO]T1 [is now holding its biennial meetings in Paris to devise its next projects]R1.

57. [Mr. Mayor ’s hope that {[references to “ press freedom ”]T1(P2) [would survive unamended]R1(P2)}P2]T1 [seems doomed to failure]R1 ;

58. [the current phrasing]T1 [is “ educating the public and media to avoid manipulation]R1. ”

59. [He]T1 [has n’t been able to replace the M’Bow cabal]R1.

60. [Other countries, including West Germany]T1 , [may have a hard time justifying continued membership]R1.

61. [The Babelists of the United Nations]T1 [are experts at obfuscation]R1.

62. [[The researchers]T1(SP1) [said]R1(SP1)]SP1 [they]T1 [have isolated a plant gene {[that]T1(P2) [prevents the production of pollen]R1(P2)}P2]R1.

63. [On a commercial scale]SP1 , [the sterilization of the pollen - producing male part]T1 [has only been achieved in corn and sorghum feed grains]R1.

64. [In a labor - intensive process]R1-1 , [the seed companies]T1 [cut off the tassels of each plant , making it male sterile]R1-2.

65. [They]T1 [sow a row of male - fertile plants nearby , {[which]T1(P2) [then pollinate the male - sterile plants]R1(P2)}P2]R1.

66. [The vast majority of the U.S. corn crop]T1 [now is grown from hybrid seeds produced by seed companies]R1.

67. [A similar technique]T1 [is almost impossible to apply to other crops , such as cotton , soybeans and rice]R1.

68. [[Mr. Leemans]T1(SP1) [said]R1(SP1)]SP1 [this genetic manipulation]T1 [does n't hurt the growth of that plant]R1.

69. [They]T1 [[attached a second gene , for herbicide resistance]T1(R1) , [to the pollen - inhibiting gene]R1(R1)]R1.

70. [Both genes]T1 [are then inserted into a few greenhouse plants , {[which]T1(P2) [are then pollinated and allowed to mature and produce seed]R1(P2)}P2]R1.

71. [One technique developed by some of these companies]T1 [involves a chemical spray supposed to kill only a plant 's pollen]R1.

72. " [There is a large market out there hungry for hybrid seeds]R1 ," [[he]T1(SP1) [said]R1(SP1)]SP1.

73. [Nevertheless]SP1 , [[he]T1(SP2) [said]R1(SP2)]SP2 , [he]T1 [is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton]R1.

74. [Already]SP1 , [the consequences]T1 [are being felt by other players in the financial markets – even governments]R1.

75. [What [triggered]R1(T1) [the latest clash]T1(T1)]T1 [was a skirmish over the timing of a New Zealand government bond issue]R1.

76. [The dispute]T1 [shows clearly the global power of Japan 's financial titans]R1.

77. [Aside from Nomura 's injured pride]SP1 , [the biggest victim]T1 [so far has been the New Zealand government]R1.

78. [[New Zealand 's finance minister]A1 , [David Caygill]A2]T1 , [lashed out at such suggestions]R1.

79. " [It]T1 [may very well be]R1 [{[what]R1-1(P2) [the Japanese banks]T1(P2) [want]R1-2(P2)}P2]R1 , " [[he]T1(SP1) [told Radio New Zealand]R1(SP1)]SP1.

80. {[Both sides]T1(P2) [are jealously guarding their turf]R1(P2)}P2 , {[and]SP1(P3) [relations]T1(P3) [have been at a flashpoint for months]R1(P3)}P3.

81. [The banks]T1 [badly want to break into all aspects of the securities business]R1

82. [And]SP1 [their suspicions of each other]T1 [run deep]R1.

83. [In the past year]SP1 , [both]T1 [have tried to stretch the limits of their businesses]R1.

84. [Mr. Conlon]T1 [was executive vice president and director of the equity division of the international division of Nikko Securities Co]R1.

85. [[As]SP1(SP1)[Yogi Berra]T1(SP1) [might say]R1(SP1)]SP1 , [it]T1 ['s deja vu all over again]R1.

86. [" Old - time kiddies , "]R1 [[he]T1(SP1) [says]R1(SP1)]SP1.

87. [[But]SP1(SP1) for the next few months]SP1 , [these boys of summers long past]T1 [are going to be reveling in an Indian summer of the soul]R1.

88. [Now that {[the baseball season]T1(P2) [is officially over]R1(P2)}P2]SP1 , [you see]SP2 , [it 's time for a new season to begin]R1.

89. " [Someone always makes you quit]R1 , " [[says]T1(SP1) [[legendary St. Louis Cardinals centerfielder Curt Flood]A1 , [the league 's commissioner]A2]R1(SP1)]SP1.

90. " [You]T1 [feel you want one more – one more at - bat , one more hit , one more game]R1. "

91. [{[When]SP1(P2) [he]T1(P2) [sent letters {[offering 1,250 retired major leaguers]T1(P3) [the chance of another season]R1(P3)}P3]R1(P2)}P2]T1 , [{[730]T1(P3) [responded]R1(P3)}P3]R1.

92. [For some players]T1 , [[the lure]T1(R1) [is money – up to $15,000 a month]R1(R1)]R1.

93. [Others , just released from the majors]T1 , [hope {[the senior league]T1(P2) [will be their bridge back into the big – time]R1(P2)}P2]R1.

94. ({[No one]T1(P2) [has worked out the players ' average age]R1(P2)}P2 , {[but]SP1(P3) [most]T1(P3) [appear to be in their late 30s]R1(P3)}P3.)

95. " [There will be a lot of malice]R1. "

96. [Men {[who]T1(P2) [have played hard all their lives]R1(P2)}P2]T1 [are n't about to change their habits]R1 , [[he]T1(SP1) [says]R1(SP1)]SP1.

97. " [If you know how to slide]T1 , [[it]T1(R1) ['s no problem]R1(R1)]R1 , " [[he]T1(SP1) [says]R1(SP1)]SP1.

98. [After all]SP1 , [[he]T1(SP2) [says]R1(SP2)]SP2 , " [[Even to make love]T1(R1) , [[you]T1(R1(R1)) [need experience]R1(R1(R1))]R1(R1)]R1.

99. [[Stewart & Stevenson Services Inc.]T1(SP1) [said]R1(SP1)]SP1 [it]T1 [received two contracts totaling $19 million to build gas - turbine generators]R1.

100. [[Statistics Canada]T1(SP1) [said]R1(SP1)]SP1 [service - industry output in August]T1 [rose 0.4% from July]R1.

101. [Both General Motors Corp. and Ford Motor Co.]T1 [have been trying to amass 15% stakes in Jaguar]R1.

102. [Many investors]T1 [certainly believe {[a bidding war]T1(P2) [is imminent]R1(P2)}P2]R1.

103. [Such a countermove]T1 [could end Jaguar 's hopes for remaining independent and British – owned]R1.

104. {[Dow]T1(P2) [will own 60% of the venture]R1(P2)}P2 , {with [Eli Lilly]T1(P3) [holding the rest]R1(P3)}P3.

105. [The 45 - year - old Mr. Kuehn , {[who]T1(P2) [has a background in crisis management]R1(P2)}P2]T1 , [succeeds Alan D. Rubendall , 45]R1.

106. [Mr. Kuehn]T1 , [[the company]T1(SP1) [said]R1(SP1)]SP1 , [will retain the rest of the current management team]R1.

107. [The gains]T1 [also sparked buying interest in other real - estate companies]R1 , [[traders]T1(SP1) [said]R1(SP1)]SP1.

108. [The balance of short positions outstanding]T1 [fell 159.7 billion yen , to 779.8 billion yen]R1.

109. [No one wants stock on their books]R1.

Appendix C

CONSENT FORM

CONSENT FORM for Audiovisual Prosody Database (APD)

Consent By signing this form, I understand and consent that my personal data, including the data categories set out below (collectively, my "Personal Data"), will be collected, processed and used within the Maria de Maeztu Strategic Programme[1] of the Department of Information and Communication Technologies (DTIC) at UPF for the purposes indicated below.

I acknowledge that I may revoke this consent at any time.

Summary In 2014 you took part in an audiovisual recording of you engaged in reading and spontaneous speech. This recording is being studied by the TALN and CMTECH research groups of the Department of Information and Communication Technologies (DTIC) Univerisitat Pompeu Fabra (UPF) within the Maria de Maeztu Strategic Programme (MdM)[2]. The purpose of this research was informed to you at the time of the recording: analyzing oral and visual prosody in connection with the communicative structure of the message.

Your voice and video data may be shared among the project members and the academic community for research in the scientific fields listed below only. Your contact information (name, e-mail, phone number) will never be revealed.

We maintain the security of the data in accordance with applicable law, and we will not make your recording or other personal data public.

By signing this form, you consent to this processing of your personal data in and relating to the recording. You may revoke your consent at any time by writing to Leo Wanner (leo.wanner@upf.edu).

[1] https://portal.upf.edu/web/mdm-dtic

[2] https://portal.upf.edu/web/mdm-dtic

Please read the text below carefully for the full details of this consent form.

Data controller The entity responsible for your data is Pompeu Fabra University (UPF) The Data Controller is the entity indicated below. All communications should be directed to the Data Protection Office or Contact Person indicated below:

- Data Controller: Universitat Pompeu Fabra, domiciled at: Plaça de la Mercé, 10-12; 08002 Barcelona (Spain), with NIF Q-5850017-D.
- Contact Office (for general data processing information): Data Protection Office / Gerencia UPF.
- Contact Person: Dr. Leo Wanner, Fax. (+34) 93 542 20 02, mail: leo.wanner@upf.edu

I note that the Universitat Pompeu Fabra may designate other contact persons and licensee Data Controllers and will indicate to me their contact person.

Background and Purposes Your data is collected and processed for the purposes stated here The Maria de Maeztu Strategic Research Program (MdM)[3] of the Department of Information and Communication Technologies (DTIC) at UPF is a research focused on data-driven knowledge extraction, boosting synergistic research initiatives across our different research areas: (1) cognitive and intelligent systems, (2) audiovisual technologies, (3) networks and communications, and (4) computational biomedicine.

The goal of the Audivisual Prosody Database ("APD") collected at the Universidad Pompeu Fabra (hence, APD-UPF), which includes my Personal Data, is to serve as benchmark data for developing adapted human-machine interaction technologies for the goals of the project, including to analyze the communicative structure correlated to prosody and facial expression, and in addition foster data sharing and reproducibility practices within the scientific community (jointly, the "Purposes").

Collected Data The data we collect includes the data indicated here as *My Personal Data* collected within the scope of the APD project, and consists of:

Duration Your data will be used for the period of the MdM project and further use My Personal Data will be used for the Purposes indicated above and any other purpose expressly authorized in writing by me. The data will be processed for the duration of the MdM project and the duration of any further scientific research uses compatible with the foregoing Purposes (including archiving, historical and

[3] https://portal.upf.edu/web/mdm-dtic

Tick	Types of Data
x	Biometric data (video/audio recording of body, face and voice)
x	Contact information (name and contact details)
x	Other personal data (age, profession, gender, language proficiency – eg. low/medium/native)

statistical purposes) in accordance with this Consent Form. I am aware that this may be perpetual, subject to my rights of revocation and deletion set out below.

Data transfers and sharing with the scientific community Your data will be used by and shared among UPF I understand that the UPF will use and shared among its researchers and other members my Personal Data for the Purposes.

APD-UPF may also be made available for research purposes within the scientific community. Whenever possible, this data will be pseudonymised prior to any further processing. Your data may also be shared among other scientific research bodies for the purposes stated above The APD-UPF may also be made available to licensees worldwide ("Licensees"), under restrictive licensing conditions and in compliance with European and local laws and regulations on personal data protection, for academic and scientific research. These Licensees will have to respect strict license conditions for the use of the APD-UPF, stipulated in the license agreement to be signed with the UPF. Licenses to access and process the Personal Data of the APD-UPF will only be granted for academic and scientific research purposes in the field of multimodal communication Including the subdomains listed above the APD-UPF will never be licensed for research purposes on any topic not listed above, and, in particular, on any topic related to person identification.

You expressly authorize this data sharing.

Whenever my Personal Data is accessed and processed by Licensees, my name and contact details will be kept and maintained confidential by the Data Controller and will not be revealed to Licensees. In this way, unless my identity is revealed through the image/voice data themselves, my identity will not be revealed to the Licensees. At the time of giving consent:

- I expressly authorise the transfer of my Personal Data to the list of entities attached in Annex hereto (if any) for processing in accordance with the Purposes and further purposes compatible with these; and
- I accept that further licensees may be interested in accessing and processing my Personal Data for these or further purposes, on a pseudonymous or anonymous basis when possible. I will be notified of any such further li-

censees and purposes, at my address set out below, in particular if the data is to be processed outside the European Economic Area.

International (non-EEA) transfers We may transfer your data outside the EU, but only in compliance with the law and your further consent Licensees may be located and established in countries outside the European Economic Area which do not have legislation similar to European data protection laws and provide adequate levels of protection (in accordance with these EU data protection laws). Nevertheless, these licensees will have to respect strict license conditions equivalent to requirements of the European Union and Spanish law/s for data protection for the use of the APD.

I understand and agree that my biometric data and data and place of birth only (excluding other personal data indicated above), as included in the APD, may be sent exclusively under such conditions to such licensees in such third countries. However, in any event my written consent will be required prior to any such transfer.

Publication We will make scientific publications about the project, but these will not include your personal data. I understand that if any results based on my Personal Data are published, my name will not be revealed, my year and place of birth may be published but my name will not be revealed, and my face image will not be published or displayed by any means without my consent (unless I agree to do so below in this Consent Form).

Security measures We apply industry standard security measures to protect confidentiality The security measures and personal data protection schemes required by law will be adopted and maintained by the Pompeu Fabra University and subsequent licensee data controllers. These measures are in accordance with the guidelines set by the National Data Protection Commissioners[4].

Data Subject Rights You may contact the original Data Controller. to exercise your rights to access, modify your data, or oppose further processing I have the right to request access to my Personal Data, and to require the Data Controller to correct and/or to delete my Personal Data, or object to any or all further processing, in conformity with applicable legislation. For these purposes, I can contact:

- The original Data Controller, the Pompeu Fabra University, at the address set out above

[4]http://ec.europa.eu/justice_home/fsj/privacy/nationalcomm/index_en.htm

- The further licensee Controllers, at the addresses notified to me when I provide my consent.

If this right is exercised, any copy of the Personal Data will be corrected, rectified and/or deleted, both in the individual APD-UPF and in any other licensed copies of the APD. For that purpose, the license agreements will specify and enforce revocation schemes.

If you have feel that your rights are violated, you may contact the Agency set out here I am informed that the data protection authority to which I may make any complaint is the Autoritat Catalana de Protecció de Dades (Catalan Data Protection Agency), that I may contact at: apdcat@gencat.cat or in writing to: Catalan Data Protection Authority, C/ Rosselló, 214, Esc. A, 1-1, 08008 Barcelona, Spain.

I confirm that I have read and understood the information sheet for the above data collection and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason. I have read the above and I understand that I am free to indicate whether or not I agree with the processing of my biometric and other personal data as described above. By signing the present form, I agree with the above stated.

Please, tick the following option if you give your consent:
I agree that my image and/or audiovisual data can be used in scientific publications and/or presentations.

Please, tick the following option if you give your consent:
I agree that my data (as identified herein) will be shared with licensees as set out above.

- Date:
- Name:
- Identification Number:
- Signature: