The Information Structure–Prosody Interface in Text-to-Speech Technologies. An Empirical Perspective

Mónica Domínguez[1], Mireia Farrús[1*] , Leo Wanner[2,1]

[1] *Universitat Pompeu Fabra, Barcelona, Spain*
[2] *Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain*

Abstract

The correspondence between the communicative intention of a speaker in terms of Information Structure and the way this speaker reflects communicative aspects by means of prosody have been a fruitful field of study in Linguistics. However, text-to-speech applications still lack the variability and richness found in human speech in terms of how humans display their communication skills. Some attempts were made in the past to model one aspect of Information Structure, namely *thematicity* for its application to intonation generation in text-to-speech technologies. Yet these applications suffer from two limitations: (i) they draw upon a small number of made-up simple question-answer pairs rather than on real (spoken or written) corpus material; and (ii) they do not explore whether any other interpretation would better suit a wider range of textual genres beyond dialogues. In this paper, two different interpretations of thematicity in the field of speech technologies are examined: the state-of-art binary (and flat) theme-rheme, and the hierarchical thematicity defined by Igor Mel'čuk within the Meaning-Text Theory. The outcome of the experiments on a corpus of native speakers of US English suggests that the latter interpretation of thematicity has a versatile implementation potential for text-to-speech applications of the *Information Structure-–prosody* interface.

Keywords: information structure, theme, rheme, prosody

Email address: monica.dominguez@upf.edu, mireia.farrus@upf.edu, leo.wanner@upf.edu (Mónica Domínguez[1], Mireia Farrús[1*] , Leo Wanner[2,1])

* Currently at the Universitat de Barcelona

1. Introduction

Natural and contextualized prosody is a key feature of synthesized speech that text-to-speech (TTS) applications have been striving to achieve since the early 1990s. But when is prosody "natural and contextualized"? According to major linguistic theories, such as, e.g., *Functional Linguistics* Halliday (1967), prosody forms part of the grammatical model of a language. Levelt (1993) and Chomsky (1995), among others, also argue that prosody renders the syntactic structure. However, this is certainly only half of the story. It is indisputable that prosody can be considered as natural and contextualized only if it reflects the communicative and emotional intentions of the speaker: what parts of the message the speaker aims to convey about what, what parts they intend to underline as central for the comprehension of the message, or how they acknowledge the reaction of their conversation counterpart (Syrdal & Kim, 2008; Wolff & Brechmann, 2015; Izzad et al., 2016; Levitan et al., 2016). Only then will the speech generated by a spoken language human-computer interface be fully adaptable to a wide range of contexts and human listeners, especially in the case of assisting conversational agents that interact with children (López-Mencía et al., 2013; Pérez-Marín & Pascual-Nieto, 2013), elderly (Wanner et al., 2017; Ortiz et al., 2007), or cognitively or mentally impaired (Wargnier et al., 2016). Moreover, perception experiments confirm that, in the case of read spoken-language transcripts, comprehension is positively affected if prosody reflects the communicative intention of the speaker (Clark & Haviland, 1977; Meurers et al., 2011; Vanrell et al., 2013). Therefore, achieving a communicatively and emotionally expressive speech prosody is decisive for the comprehension of the synthesized message.

The communicative intention of the speaker of an utterance is captured by a number of linguistic theories in terms of the *Information Structure* (Lambrecht, 1994; Erteschik-Shir, 2007; Krifka, 2008), the *Topic-Focus articulation structure* (Hajičova et al., 1998; Sgall et al., 1973), or the *Communicative Structure*

(Mel'čuk, 2001). The most common of them is the Information Structure (IS). In accordance with the multidimensionality of the communicative intention of a speaker (cf. above), IS is multidimensional, which is also acknowledged in several theoretical studies that investigate the relation of IS to prosody; see, among others, (Calhoun, 2010; Baumann, 2012; Vallduví, 2016). However, in the context of speech technologies (in particular, TTS applications), the IS in the "IS–prosody interface" is usually reduced to a single dimension that signals what parts of the message the speaker aims to convey about what (or, "what the utterance is about" and "what is being uttered about it"). This dimension is also referred to as *theme-rheme* (or *thematicity*) structure Mathesius (1929); Halliday (1967); Steedman (2000); Mel'čuk (2001).[3]

In our work, we focus on the thematicity dimension. Studies that have been carried out on the IS–prosody interface in the context of speech technologies are limited to the rather simple correlation of a flat binary theme–rheme structure with rising–falling intonation contours in terms of the labels from the ToBI[4] convention (Silverman et al., 1992) on short made-up question–answer examples; see, e.g. (Büring, 2003; Steedman, 2000). An implementation on the grounds of Steedman (2000)'s characterization was tested in TTS applications by (Kruijff-Korbayová et al., 2003; Haji-Abdolhosseini, 2003; Kügler et al., 2012). However, neither a simple binary partition of a sentence independently of the complexity of a sentence into theme and rheme nor a question-answer setting is apt to cope with the problem of monotonous prosody in TTS applications.

In order to provide a sufficiently fine-grained and comprehensive foundation for a computational communicative model of prosody to be implemented in

[3] The first theoretical studies on theme–rheme go back to the Prague school's founder Mathesius (1929). Later, the Prague school adopted the term *topic–focus* instead (Daneš, 1970; Hajičova, 1987) to refer to this concept, as other authors from different schools of linguistics do (Von Stechow, 1981; Rooth, 1992; Lambrecht, 1994). A number of other studies refer to it as *Givenness* (Schwarzschild, 1999), and thus talk about *given–new* information (Chafe, 1976; Clark & Haviland, 1977; Brown, 1983).

[4] "ToBI" stands for ***To**nes and **B**reak **I**ndexes*.

speech technologies, such that once the communicative intention of the speaker (or a conversational agent) is given in terms of a formal representation of thematicity, adequate and varied prosody is produced, the following objectives need to be addressed:

(i) to identify a formal representation of the thematicity structure that can be annotated given any type of text format (dialogue or monologue) and that can serve for the derivation of prosody;

(ii) to carry out empirical studies on the correlation of the thematicity structure with prosody on a syntactically varied sentence collection.

This is not to say that prosody does not depend on other layers of the linguistic description such as syntax and phonology (Selkirk, 1984), or, in particular, on pragmatics (Hirschberg, 2008). Especially pragmatics plays an outstanding role. However, since, on the one side, syntax and phonology are also influenced by thematicity and, on the other side, the formal modeling of contextual features and their relation to prosody for TTS applications from the perspective of pragmatics is currently beyond the reach in the field, the undisputed prominence of thematicity for prosody justifies a focus on the structural linguistic description and thus on (i) and (ii). In our work, we address (i) by exploring a formal tripartite hierarchical representation of thematicity as proposed by Mel'čuk (2001) in the context of the Meaning-Text Theory (MTT), which proved to be instrumental for such Natural Language Processing (NLP) tasks as Natural Language Generation (NLG) (Bouayad-Agha et al., 2012; Ballesteros et al., 2015); and (ii) by investigating empirically whether the correlation between the representations of thematicity and prosody is bidirectional, i.e., whether prosody can be derived from thematicity and vice versa. For this purpose, we carry out machine learning-based classification experiments on a spoken language corpus,[5] which

[5] TTS applications make use of machine learning algorithms for several tasks, including prosody prediction. That is why it is essential that our empirical analysis is devised following a machine learning methodology.

consists of an extract of 109 isolated sentences from the popular Wall Street Journal (WSJ) corpus (Charniak et al., 2000), read aloud by native speakers of English. We opted for a reading-aloud setup because one of our applications is a "reading aloud" agent (Domínguez et al., 2018), and deficiencies in expressive prosody in TTS become evident with the syntactically demanding genre of newspaper material.

The sentences in our corpus are annotated with their thematicity structure (both MTT's tripartite hierarchical thematicity and the flat binary theme–rheme dichotomy, which constitutes the state of the art in speech technologies and which we use as the reference thematicity structure) and with their prosodic structure (in terms of acoustic parameter-oriented labels automatically derived from three prosodic elements, namely, F0, intensity and rhythm, and in terms of ToBI labels). First, we assess to what extent a standard machine learning model is able to predict from the given thematic features of both thematicity structures of a sentence the ToBI label of each word in this sentence. We observe that the tripartite hierarchical thematicity structure renders a considerably higher performance of the model – which leads us to assume that the tripartite hierarchical thematicity structure correlates better with intonation than a bipartite thematicity structure does. Then, we explore whether a machine learning model can predict from the acoustic parameters of a word the label of the thematic span to which this word belongs. The outcome of this experiment shows that, indeed, different thematicity spans have distinct prosodic characteristics.

To the best of our knowledge, the presented study is the first attempt to prove that linguistically-motivated speech technologies can be modelled on empirical grounds, apart from a broad characterization of theme and rheme containing rising and falling patterns that is based on Steedman (2000)'s work and which has been tested in content-to-speech (CTS) applications by, e.g., Kruijff-Korbayová et al. (2003); Haji-Abdolhosseini (2003); Kügler et al. (2012).

The next section introduces the fundamentals of thematicity and prosody as used in our work. Then, in Section 3, the experimental setup and the results of the experiments to predict, on the one side, intonation patterns using thematicity, and, on the other side, thematicity using prosody patterns are outlined. Section 4 analyzes the outcome of these classification experiments. Section 5, finally, draws some conclusions and discusses the implementation potential of the IS–prosody interface on the grounds of the MTT in the light of our experiments.

2. Fundamentals

Since we focus on thematicity and its relation to prosody, we refrain, in what follows, from the review of the vast volume of research on Information Structure, Topic Focus Articulation and the Communicative Structure, and merely introduce the notions of thematicity that we work with and the correspondence of both of them with prosody.

2.1. Thematicity Structure Annotations

In this section, we introduce the flat binary thematicity structure, which we use in our experiments as reference thematicity structure used in state-of-the-art TTS applications, and the hierarchical tripartite thematicity structure, which forms part of our thematicity–prosody proposal.

2.1.1. Flat binary thematicity structure

The theme–rheme structure observed in the implementation of the IS–prosody correspondence in state-of-the-art TTS applications partitions a sentence into two subsequent spans, namely *theme* and *rheme*, such that it is a binary flat division of a sentence related to both discourse and syntax layers; see, e.g., (Erteschik-Shir, 2007; Haji-Abdolhosseini, 2003; Steedman, 2000). In order to identify these divisions, a question (Q) is constructed that shall help identify the theme and the rheme: theme is echoing the question, while rheme is the information (A) provided to answer the question. Consider, as illustration, example (1) for this interpretation of theme–rheme based on an example from our corpus ('T' stands for "theme" and 'R' for "rheme").

(1) Q: *What did he say?*

A: [*The proposed rules also would be tougher on the insiders still required to file reports*]$_{\mathrm{R}}$, [he said]$_{\mathrm{T}}$.

Note, however, that when the question cannot be unambiguously derived from the context, or when the context allows for more than one question, a different thematicity segmentation is possible:

(2) Q: *What did he say about the proposed rules?*

A: [*The proposed rules*]$_{\mathrm{T}}$ [*also would be tougher on the insiders still required to file reports*]$_{\mathrm{R}}$, [*he said*]$_{\mathrm{T}}$.

2.1.2. Hierarchical tripartite thematicity structure

A different view on the thematicity structure as presented above is advocated by I. Mel'čuk in the context of the Meaning-Text Theory (MTT) (Mel'čuk, 2001). MTT's thematicity 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 a statement, and (ii) the fact that thematicity is defined over *propositions*, 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 the theme(T1)/rheme(R1)/specifier(SP1) distribution in our example sentence in the sense of Mel'čuk compared to (1):[6]

(3) {[*The proposed rules*]$_{\mathrm{T1}}$ [*also would be tougher on the insiders still required to file reports*]$_{\mathrm{R1}}$, {[[*he*]$_{\mathrm{T1(SP1)}}$ [*said*]$_{\mathrm{R1(SP1)}}$}$_{\mathrm{P2}}$]$_{\mathrm{SP1}}$}$_{\mathrm{P1}}$.

In (3), the hierarchical thematicity structure is represented at different levels:

[6] The annotation is copied from Bohnet et al. (2013). Also note that this and all following annotations of the tripartite hierarchical thematicity structure are aimed to follow the principles and criteria put forward in (Mel'čuk, 2001). Given that theme /rheme determination criteria are still an active research topic, these principles may not be in full accordance with the principles of other works. However, punctual discrepancies concerning the identification of theme/rheme across different theories do not diminish the insights we obtain in the experiments presented in this paper.

- at level 1 (L1), the proposition P1 contains a theme, a rheme and a specifier;
- at level 2 (L2), SP1 contains the proposition P2, which has a theme T1(SP1) and a rheme R1(SP1).[7]

That is, spans at level 2 are embedded into one of the thematic elements (namely SP1) of level 1. Overall, the annotation observed in (3) is determined in accordance with the following annotation guidelines sketched in Bohnet et al. (2013):

1. in a hypotactic 'direct speech clause–reporting clause' sentence construction, the reporting clause is the Specifier of the proposition (P1), which is formed by the whole sentence;
2. the full reporting clause forms an embedded proposition (P2);
3. *The proposed rules . . .* answers to the question "What about the proposed rules?", such that *The proposed rules* is the Theme of P1;
4. *also would be tougher on the insiders still required to file reports* can be negated: *would <u>not</u> be tougher on the insiders still required to file reports* and is thus the Rheme of P1;
5. in P2, in accordance with the criteria in 3. and 4., *he* is the theme and *said* is the rheme.

In sentences that consist of coordinated propositions, as in example (4), there is a parallel thematicity structure, one partition per proposition, at level 1.[8] The two partitions are labeled as 'P2' and 'P3' respectively, with a T1/R1 division of each at level 1.

[7] In the annotation of examples such as (3), the proposition markers can be dropped for the sake of the simplicity (and thus clearness) of the annotation: P1 covers the whole sentence (and thus does not contribute any distributional information, and P2 coincides with SP1. Further details on the thematicity annotation schema are defined in Bohnet et al. (2013).

[8] An additional sentential proposition can be assumed, which contains all coordination elements. However, since this proposition does not imply an own thematicity distribution, we omit it in our example.

(4) Coordinated propositions at level 1.

	No one has worked out	the players' average age,	but	most	appear to be in their late 30s.
L1	P2		P3		
	R1	T1	SP1	T1	R1

2.2. Prosody Annotation

Prosody is usually defined in terms of three acoustic elements:[9] 1. fundamental frequency (F0) or *pitch* as its perception correlate, 2. intensity (usually perceived as loudness), and 3. rhythm (including speech rate, pauses and segment length) (Beckman & Pierrehumbert, 1986; Ladd, 2008).

Conventions in the area of Speech Prosody encode prosodic information using a symbolic label alphabet, such as INTSINT (Hirst & Di-Cristo, 1998), iViE (Grabe et al., 1998) and the ***To**nes and **B**reak **I**ndexes* (ToBI) (Silverman et al., 1992). ToBI, which is the most well-known of them, represents the intonation contour (changes in F0) by means of discrete labels indicating whether the F0 is a high (H*), a low (L*) tone or a combination of both, depending on a post- or pre-nuclear rise or fall of F0 (L*+H, H*+L, L+H*, H+L* respectively). According to ToBI guidelines, all prominent words (called *pitch accents*) must be labeled; cf. (5).[10]

(5) Marking of intonation contours by ToBI labels

	Ever since,	the remaining members	have been desperate	for the United States	to rejoin this dreadful group.
ToBI	L*+H LL%	H* LH-	L* LL%	H* LH-	H* L* LL%
Them L1	SP1	T1	R1		
Them L2				P2	
				T1	R1

For our study and experiments, we use the ToBI convention and a parametric representation of prosody. Since the capabilities of modifying the prosodic contour by means of ToBI labels by speech synthesizers such as Festival (Black & Taylor, 1997) and MaryTTS (Schröder & Trouvain, 2003) are rather limited

[9] Some authors follow Campbell's proposal (Campbell & Mokhtari, 2003) to include voice quality as a fourth element of prosody.

[10] Syllable nuclei are referred to in ToBI by the '*' symbol.

to slight changes in the amount of increase in F0 (e.g., +50% increase when inserting a H* symbol or a slight modification of the end of a word when applying an LH%), we annotate the corpus with a reduced catalog of ToBI labels that can actually be mapped to perceivable modifications in the TTS engine. Table 1 shows the inventory of ToBI labels used in the annotation. Words that are prosodically unmarked ('False') and words that carry a prosodic label ('True') are annotated in each prosodic phrase. 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 1.

Table 1: Catalog of ToBI labels

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

A parametric representation of prosody allows us to not only analyze F0 variations, but also to introduce alternative prosodic cues, such as, for instance, speech rate and intensity variation, and even a combination of them, which ensures a wider range of variability of the synthesized speech. Table 2 shows the complete list of the twelve acoustic parameters (grouped by the three acoustic elements F0, intensity, and rhythm) used in our experiments within the parametric representation. To extract the absolute values, we use an extension of the Praat software for feature annotation (Domínguez et al., 2016).[11] Normalized values relative to the whole sample (z.score) and to the previous span

[11] For the original Praat software, see Boersma (2001).

(z-score_prev.sp) at the same level of embeddedness (if applicable) are computed for each thematicity span taking mean values of each prosodic element. Parameters that refer to a time point ('maxF0.t' and 'minInt.t') are computed extracting the point of maximum F0 and minimum intensity respectively and calculating the relative position in the span with a minmax score. In other words, the computed score provides information on the location of the F0 peak and intensity valley. Thus, if an F0 peak is located at the beginning of the span, it will have a score between 0 and 0.5, and if an intensity valley is located at the very end of the span, the score will be close to 1.

Table 2: Prosodic elements and acoustic parameters used in our experiments

	F0	Intensity	Rhythm
Absolute	mean F0 (in Hz)	mean intensity (in dB)	dur (in sec), speech rate (in words/sec)
Relative	z-score_F0 (z_F0) z-score_F0_prev.sp maxF0.t	z-score_intens (z_int) z-score_intens_prev.sp minIntens.t	z-score_speech.rate (z_sr) z-score_dur_prev.sp

2.3. Correlation between thematicity and prosody: An illustration

In this section, we exemplify with sentences from our corpus the correspondence between thematicity and prosody that we are going to explore in our machine learning experiments. Consider the following spoken sentence from our corpus:

(7) *Ever since, the remaining members have been desperate for the United States to rejoin this dreadful group.*

Example (7a) illustrates its binary flat theme–rheme division as considered in a state-of-the-art CTS application; cf. (Kruijff-Korbayová et al., 2003; Haji-Abdolhosseini, 2003; Kügler et al., 2012).

(7a) 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

(7b) shows its segmentation according to the hierarchical tripartite thematicity structure: 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)).

(7b) [*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 1 shows the fundamental frequency and intensity contours as spoken by one of the participants in our experiments and both thematicity annotation schemes: (a) displays the binary flat thematicity, and (b) displays the tripartite hierarchical thematicity.

The triangles and bullets in Figure 1 capture the mean acoustic parameters (F0 represented as triangles over the F0 contour line in the upper part of the figure, and intensity represented as bullets over the intensity line below) across (a) the binary flat theme–rheme structure and (b) the tripartite hierarchical thematicity structure. As can be observed in (a), the scores that originate from the flat thematic structure result in linear functions that do not reflect any marked distinction between theme and rheme. The tripartite division of thematicity already reflects more variability in terms of prosodic parameters. Moreover, as R1 is further subdivided into T1 and R1 at L2, these L2 divisions not only add a richer prosodic characterization of (7), but also suggest a clear distinction between theme and the rest of spans in terms of acoustic parameters, as seen in (b). A tripartite hierarchical segmentation of thematicity is hypothesized to better capture the acoustic variability of this speech sample. This hypothesis is further tested in the next section in an experiment for predicting ToBI labels using both the state-of-the-art binary flat thematicity and the hierarchical thematicity model. Even though it may initially seem a “straw man” comparison,

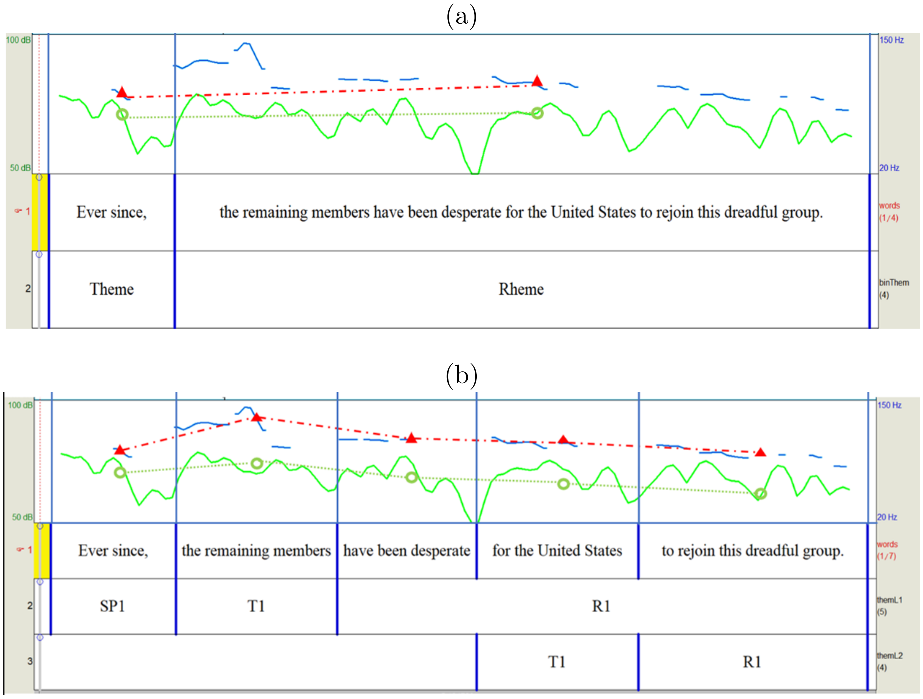

Figure 1: Illustration of F0 and intensity curves, and segmentation into binary (a) and hierarchical (b) thematicity of (7)

our intention is to prove the concept in an experimental setting before setting out to annotate a large amount of text embarking on a time-costly implementation in a TTS architecture, which may lead to little improvement of existing simpler strategies, such as a by-default F0 decay in statements.

3. Experimental study setup

As already mentioned in Section 1, we aim to assess in classification (or supervised machine learning) experiments on an extract of 109 isolated sentences of the WSJ corpus (Charniak et al., 2000) the potential of thematic features to predict intonation and the potential of prosodic features to predict thematic features. In what follows, we first provide some further details on the corpus and outline then the setup and results of the experiments. The results are then discussed in a separate section (Section 4).

3.1. Corpus of read speech used in the study

For our experiments, the textual and speech characteristics of the 109 WSJ sentences that we work with are of relevance. Let us thus have a closer look at both.

3.1.1. Textual characteristics of the experimental corpus

The 109 sentences were selected to cover a rich variety of syntactic constructions, ranging from simple clauses to coordination, subordination and the combination of both. This variety ensures a minimum amount of thematicity in terms of: (i) the number of hierarchical levels of thematicity (up to three in this corpus), (ii) the presence/absence of each type of thematicity span, (iii) the position of spans within the sentence and with respect to each other, and (iv) the continuity of spans or lack of it. Given that punctuation is known to affect prosodic phrasing when reading (Kalbertodt et al., 2015), a representative number of punctuation marks was also taken into account for the selection, including question marks, quotes, semi-colons and commas with different functions. In average, a sentence of our corpus contains fifteen words, with a minimum of three words and a maximum of thirty. 54% of the sentences contain between sixteen and twenty-three words.

From the perspective of the binary flat theme–rheme structure, 98% of the sentences contain a theme and a rheme, while 2% are all-rhematic. From the perspective of the hierarchical tripartite thematicity structure, all sentences contain a rheme (R1), 98% have a theme (T1), and 30% include a specifier (SP1). One-word themes represent 31% of the total number of themes (there are 152 themes in the dataset, including all levels of embeddedness), and 44% contain more than three words. 70% of the sentences consist of one proposition (P1) with thematicity partitions (T1, R1 and SP1) only at level 1; 14% of the sentences contain more than one proposition (coordinated or subordinated P2, P3, and so on) split into thematicity spans at a deeper level (e.g., T1(P2) and R1(P2)); and 16% of the sentences involve spans embedded in other thematicity spans. In this last group, most thematicity spans that are split into level 2 elements are specifiers (68%) (e.g., T1(SP1) and R1(SP1)), followed by rhemes (24%) and themes (8%).

The binary flat theme–rheme annotation has been carried out and consensualized by two linguists working on the topic of Information Structure; the hierarchical tripartite thematicity annotation stems from Bohnet et al. (2013).

3.1.2. Speech characteristics of the experimental corpus

To produce the read-aloud pendant of the 109 sentences, twelve native speakers of American English born in different dialectal regions in the USA were recruited for a recording session in a professional studio. Six of them were male and six female; their age ranged between 20 and 61 years. Such a wide variety of speaking styles was expected to precisely raise questions on firm hypotheses about the IS–prosody correspondence. For instance, do rising tunes (i.e., L*+H LH% ToBI labels) occur in theme spans regardless the dialectal origin, gender or age of the speakers? Speakers were asked to read the whole set of sentences naturally, making a short pause of approximately three seconds after each sentence since our study targets the sentence as referent linguistic unit. They were also instructed to take the initiative and repeat sentences if they felt a sentence did not sound natural, a word had been mispronounced (some sentences contained low frequency words even for a journalistic discourse), or words were grouped together awkwardly.

The word tokens of the recorded sentences were manually annotated with eight ToBI labels (see Table 1) by an expert in prosody and a subset of the annotations was validated by two specialists with a coincidence rate in 72% of the labels, which means that they substantially agree on the annotations. In addition, the thematic spans of the hierarchical tripartite thematicity annotation have been assigned acoustic parameters that have been computed from the pitch and intensity objects generated with Praat (see Subsection 2.2).

3.2. Outline of the Experiments

In accordance with our goal to investigate the mutual dependence between the thematic and prosodic features, we carried out two rounds of classification experiments. As already mentioned before, in the first round, we aim to predict prosodic ToBI labels from thematic features, and in the second round, the thematicity span labels from acoustic parameter features. For the classification, we draw upon the repertoire of classifiers from the Weka toolkit (Hall et al., 2009).[12]

3.2.1. Predicting ToBI labels by thematic features

The task in this round of experiments has been to assign to each word in a given sentence from our working corpus one of the ToBI labels, drawing upon a number of extracted sentential features, which are summarized in Table 3. The prediction with binary flat thematicity structure uses, in accordance with its definition, two features (one for theme and another for rheme), while

[12] Weka is an open source machine learning software package that can be accessed through a graphical user interface, standard terminal applications, or a Java API (see also `https://www.cs.waikato.ac.nz/ml/weka/`). It is widely used for teaching, research, and industrial applications, and additionally gives transparent access to well-known classification algorithms.

the prediction with hierarchical thematicity uses features that account for the description of hierarchical thematicity, namely the tripartite thematicity labels (theme, rheme and specifier) divided in two levels (L1 and L2). In order to account for the overall thematicity structure within the sentence, features that specify the span position and the total number of spans in the sentence are also used in both settings.

Table 3: Thematicity features used in the experiment for the prediction of ToBI labels by the baseline, binary flat thematicity, and hierarchical thematicity models

Type	**Feature**	**# Distinct features**	Features
General	Word Position	28	Numbers from 1 to 28
	Total n. Words	22	Numbers from 4 to 28
Thematicity	Binary Flat	2	T / R
	Tripartite Hierarchical (L1)	6	T1/R1/SP1/SP2/R1-1/R1-2
	Tripartite Hierarchical (L2)	3	T1/R1/SP1
	Total n. Spans	12	Numbers from 1 to 12
	Span Position	10	A/B/C/D/(. . .)/Z

To have a baseline with which the performance of both thematicity structures can be compared, we also assess the ToBI label prediction potential by two numeric features (referred to in Table 3 as ‘General’): (1) the number of words in the sentence in question, and (2) the relative position of the word that is to be labeled in the sentence. In all three runs (i.e., with 1. baseline features, 2. flat binary thematicity structure, and 3. hierarchical tripartite thematicity structure as predictors), the Weka J48 decision tree classifier with 10-fold cross-validation has been used.[13]

[13] Compared to other classification models (including, e.g., the nowadays more popular neural network models), decision tree classifiers have the advantage that they allow for the inspection of each decision taken by the model with respect to its plausibility, which is of great value in an empirical study like ours. Furthermore, due to the small amount of annotated text with thematicity structure gathered so far, neural network models cannot be trained

Table 4: Quality of the prediction of ToBI labels using numeric baseline features (‘BL’), binary thematicity structure features (‘$\mathrm{BF}_{\mathrm{T-R}}$’) and hierarchical tripartite thematicity structure features (‘$\mathrm{TH}_{\mathrm{T-R}}$’) in terms of Precision, Recall, and F1-measure on the complex sentences subcorpus of 52 sentences

	Precision			Recall			F-measure		
	BL	$\mathrm{BF}_{\mathrm{T-R}}$	$\mathrm{TH}_{\mathrm{T-R}}$	BL	$\mathrm{BF}_{\mathrm{T-R}}$	$\mathrm{TH}_{\mathrm{T-R}}$	BL	$\mathrm{BF}_{\mathrm{T-R}}$	$\mathrm{TH}_{\mathrm{T-R}}$
S	0.45	0.52	**0.64**	0.28	0.40	**0.64**	0.35	0.46	**0.64**
L*+H	0.47	0.55	**0.74**	0.64	0.73	**0.84**	0.54	0.62	**0.78**
LL%	**0.50**	0.40	0.43	0.03	0.16	**0.44**	0.05	0.23	**0.43**
LH%	0.53	0.65	**0.87**	0.75	0.80	**0.94**	0.62	0.72	**0.90**
U	0.42	0.45	**0.57**	0.08	0.16	**0.37**	0.14	0.24	**0.45**
H*	**0.86**	0.86	0.81	0.71	0.78	**0.82**	0.78	0.81	**0.82**
L*	0.52	0.49	**0.58**	0.11	0.25	**0.44**	0.18	0.33	**0.50**
HL%	0.26	0.40	**0.51**	0.02	0.10	**0.23**	0.05	0.16	**0.32**
Average	0.51	0.58	**0.73**	0.52	0.61	**0.75**	0.48	0.57	**0.74**

Given that for simple clause sentences both types of thematicity structures result in the same theme/rheme annotation (and thus in the same ToBI labels), and we are interested in their contrastive comparison from the perspective of intonation pattern prediction, we separately execute the models on the 52 (out of 109) syntactically complex sentences (i.e., those sentences that contain at least two clauses and/or temporal or spatial circumstantials and for which both thematicity structures differ) of our annotated corpus and on the entire corpus of 109 sentences. Table 4 displays the results of the experiment on the 52 complex sentences in terms of the metrics Precision, Recall, and F1-measure,[14] while Table 5 displays the average precision figures from the experiment on the entire set of 109 sentences. We opt for showing only the average precision scores for this latter experiment run because Table 4 already displays in detail

sufficiently well to compete with classical machine learning models.

[14] Precision is the ratio of correctly assigned ToBI labels from all labels that have been assigned; Recall is the ratio of correctly assigned ToBI labels from all labels that had to be assigned, and the F1-measure is the harmonic mean of Precision and Recall.

the superiority of the hierarchical tripartite thematicity model, and we merely aim to assess how the overall prediction quality changes in a realistic setup in which complex and simple sentences are intermingled.

Table 5: Precision scores comparison between models executed on the entire corpus of 109 sentences (ALL) and on the reduced corpus of 52 complex sentences (RED)

	BL	$\mathrm{BF}_{\mathrm{T-R}}$ model	$\mathrm{TH}_{\mathrm{T-R}}$
P(ALL)	0.47	0.50	**0.65**
P(RED)	0.51	0.58	**0.73**

3.2.2. Predicting thematicity labels by acoustic parameter distributions

Having shown the appropriateness of the tripartite hierarchical thematicity structure for the generation of ToBI labels at the word level, and its superiority compared to the binary flat structure, we continue to work with the $\mathrm{TH}_{\mathrm{T-R}}$ structure only. In what follows, we aim to show that acoustic parameters can identify labels of thematicity spans at all levels of the thematic hierarchy, i.e., that the correlation between hierarchical thematicity and prosody is bidirectional. We assume a thematicity span to be determined by its category (theme, rheme, or specifier) and its embeddedness into a concrete thematic category. ‘R1’, ‘R1(SP1)’, ‘R1(T1)’ would be thus examples of communicative spans. The features used for the classification are the aforementioned acoustic parameters (see Table 2).[15]

The experiment consists in predicting the label of a given hierarchical thematic span based on acoustic parameters and the number of words in this span. For this purpose, we train as classifier (or predictor) a bagging classifier, again, from the Weka toolkit on the acoustic parameters and $\mathrm{TH}_{\mathrm{T-R}}$ structures annotation of our 109 sentences corpus. As baseline, which serves to evaluate and

[15] To indicate the position of a span, A and Z are used for the first and last segment respectively, and consecutive letters in alphabetical order are optional and depend on the total number of spans in each sentence.

compare the level of improvement in the light of the unbalanced nature of our corpus, we use the ZeroR classifier, which predicts the majority class.

Table 6 shows the average precision (P), recall (R) and F-measure (F1) for each thematicity label across the speech of all 12 speakers for the baseline (BL) and the bagging classifier (BC).

4. Discussion of the outcome of the experiments

In this section, we assess the outcome of the experiments outlined in Section 3.2 with respect to the potential of the two different thematicity structures to predict intonation labels and the possibility to derive (hierarchical) thematicity elements from an acoustic parameter distribution.

4.1. Assessing the potential of thematicity features to predict ToBI labels

Let us first analyze in some detail the performance of the different models on the complex sentences subcorpus. Table 4 above shows that the tripartite hierarchical thematicity structure yields an improvement of the proposed IS–prosody interface over the baseline and binary flat thematicity. F-measure of L*+H and LH% increases with the hierarchical thematicity by 0.16 points and 0.18 points respectively compared to the binary flat representation and by nearly 0.20 points compared to the baseline. These results suggest that the hierarchical thematicity is able to generate correctly more prosodic variation by means of bitonals, especially for rising pitch accents (PA) and boundary tones (BT). In particular, as the precision of 0.87 and the recall of 0.94 show, LH% is accurately captured, much better than both BL and $\mathrm{BF}_{\mathrm{T-R}}$ do. There is only one ToBI label for which BL and $\mathrm{BF}_{\mathrm{T-R}}$ achieve a higher precision (0.86) than $\mathrm{TH}_{\mathrm{T-R}}$, namely H*. However, the recall is for $\mathrm{TH}_{\mathrm{T-R}}$ considerably higher, such that the F1-score is also higher. Among the less well predicted ToBI labels are U (for lexically unstressed words) and the boundary tones LL% and HL%.

We show in Table 7 and Table 8 the confusion matrices of the two models that predict ToBI labels based on binary flat thematicity (cf. Table 7) and

Table 6: Average performance of the baseline (BL) and the bagging classifier (BC) for predicting hierarchical thematicity labels, based on acoustic parameters

	Precision		Recall		F-measure	
	BL	BC	BL	BC	BL	BC
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**

hierarchical tripartite structure (cf. Table 8). The analysis of these confusion matrices provides evidence that the model that draws upon hierarchical thematicity performs better in the prediction of pitch accents and boundary tones, in particular of those related to rising tunes for complex long sentences.

The confusion matrix of the hierarchical tripartite structure model in Table 8 shows L*+H is confused with H* or L* labels, that is, with other pitch accent labels, but never with boundary tones (i.e., LL%, LH%, HL%). Boundary tones are mostly confused among each other, that is, LH% with HL% and LL%, and to a lesser extent with categories with the highest number of instances (S and U). The confusion matrix of the flat binary structure model in Table 7 shows errors across all categories; for instance, L*+H is confused, apart from H* and L*, with S and U labels and even with the boundary tone LH%. It has also the tendency to produce more errors with S and U categories than with boundary tones.

Table 7: Confusion matrix: Prediction of ToBI labels by the model based on the features of the flat binary thematicity structure

a	b	c	d	e	f	g	h	← classified as/↓ is in reality
376	235	1	267	32	1	8	9	a = L*+H
77	**1879**	6	528	63	5	7	15	b = S
17	81	**43**	44	4	26	38	8	c = LH%
85	483	0	**2539**	35	1	9	1	d = U
109	391	2	246	**152**	2	19	24	e = H*
14	68	20	58	4	**654**	21	3	f = LL%
11	95	29	84	11	75	**105**	5	g = HL%
29	206	6	110	38	0	6	**44**	h = L*

As S and U are the categories with the highest number of instances, the confusion matrices of both models reflect a bias towards these two labels. In an implementation setting, this bias could be addressed by introducing a rule of lexical stress assignation (as most TTS applications in fact do), that is, a rule that assigns lexical stress to content words but not to function words (e.g., pronouns, determiners, articles, etc.). Apart from the S/U bias, confusion matrices show that bitonal pitch accents (L*+H) tend to be confused with the H* label, but not with boundary tones (those signaling the end of an intonational phrase).

Table 8: Confusion matrix: Prediction of ToBI labels by the model based on the features of the hierarchical tripartite thematicity structure

a	b	c	d	e	f	g	h	← classified as/↓ is in reality
598	148	0	87	76	3	1	16	a = L*+H
95	**2167**	24	154	88	7	15	30	b = S
1	30	**115**	7	4	41	62	1	c = LH%
26	143	0	**2968**	13	0	0	3	d = U
169	247	7	109	**353**	11	4	45	e = H*
1	22	38	29	8	**694**	48	2	f = LL%
0	24	75	32	3	95	**184**	2	g = HL%
50	159	9	34	76	6	2	**103**	h = L*

Results from the experiment run on all 109 sentences of our corpus (i.e., including both simple and complex sentences), prove that even when instances include a large amount of simple sentences (nearly 50% of the whole corpus), the difference in precision scores between the binary flat thematicity representation ($\mathrm{BF}_{\mathrm{T-R}}$) and the hierarchical tripartite thematicity model ($\mathrm{TH}_{\mathrm{T-R}}$) are nearly the same (i.e., around -0.150 points in favor of $\mathrm{TH}_{\mathrm{T-R}}$). Both $\mathrm{TH}_{\mathrm{T-R}}$ and $\mathrm{BF}_{\mathrm{T-R}}$ outperform the baseline; $\mathrm{TH}_{\mathrm{T-R}}$, again, with a higher margin than $\mathrm{BF}_{\mathrm{T-R}}$. It is interesting to note that the precision scores decrease for both $\mathrm{BF}_{\mathrm{T-R}}$ and $\mathrm{TH}_{\mathrm{T-R}}$ when executed on the entire corpus (0.8 for both). This means that prediction of ToBI labels for simple sentences is a bigger challenge in the case of simple sentences, for which the thematic features are rather limited.

Overall, our experiments on the prediction of ToBI labels by thematic features have shown that the tripartite hierarchical thematicity structure contributes to the improvement of the prediction of intonation labels compared to the binary flat theme–rheme structure, which is commonly used in state-of-the-art implementations of the IS–prosody interface in computational settings. The hierarchical thematicity structure is more accurate, especially for prediction of bitonal labels, i.e., rising pitch accent (L*+H) and rising and falling boundary tones (LH% and HL%, respectively), which are instrumental for the generation of communicative pauses and a varied range in prominent intonation in long complex sentences containing few punctuation marks. However, as the analysis of the confusion matrices shows, further exploration in this direction is needed using a larger corpus that includes several samples from different dialects, gender, and age groups. These socio-linguistic variables are well-known to affect prosody, and in our corpus we only have one speaker per dialectal region. Therefore, we cannot reach definite conclusions, even though the confusion matrices show that there might be different realisations of the same thematicity phenomena.

4.2. From acoustic parameters to thematicity

In this section, we assess the potential of acoustic parameters to predict hierarchical thematic span labels. Overall, the results of this experiment in Table 6 prove an interesting prediction potential of acoustic features for thematicity labels as described within the MTT framework by Mel'čuk (2001). If we compare the improvement over the baseline classifier, a considerable increase in all measures is obtained (Precision: +0.64, Recall: +0.49, and F1: +0.62). Moreover, a precision of $\geq$ 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 versatile representation of the Information Structure than the traditional flat binary theme–rheme.

Table 6 also reveals some interesting details. Thus, the majority class baseline is able to identify only R1 with a precision of 0.22 and a recall of 1, which is not surprising since R1 dominates the thematicity spans. Apart from R1, no other thematicity span labels are identified. In contrast, the bagging classifier has a very low recall for R1, but a precision of 1. This means that in the case of a clearly dominating thematicity span, our bagging classifier model overfits, or, in other words, is able to recognize only cases which it learned during the training procedure. For the other thematicity span labels, varying precision / recall (and thus also F1-measure) figures are achieved. The most problematic of them are R1-1 with an F1-measure of 0.21 and T1(SP2) with an F1-measure of 0.20. This is likely due to the low number of corresponding labels in the training partitions of our corpus.

Table 9 shows the confusion matrix of the bagging classifier. Different thematicity spans 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 and both are 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 phenomenon occurs in embedded rheme spans (R1(SP1), R1(R1), R1(P2), R1(P4) and R1(P5)).

5. Conclusions

Theoretical studies on the Information Structure–prosody interface have stated for some time now that there is a correlation between how the linguistic content is structured communicatively and the way intonation is used in natural speech to convey this communicative structure. While state-of-the-art implementations of the IS–prosody interface in TTS applications rely on descriptions based upon simple made-up examples, in the present study, this correlation has been investigated from the formal representation of hierarchical thematicity within the MTT framework on a corpus of read speech in American English. The study also puts forward the importance of pivoting the transition between theoretical studies and computational applications. We highlight the relevance of exploring formal representations of thematicity, such as the MTT's and we foresee promising outcomes when used as basis for implementation of

Table 9: Confusion matrix: prediction of thematicity labels by acoustic parameters

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/ ↓ is in reality
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)

communicatively-oriented models in TTS and conversational agent applications. Preliminary experiments have been carried out to implement a thematicity-to-prosody module in English and German (Domínguez et al., 2017, 2018). In this context, it should be, however, noted that these supervised classification experiments are different from an actual implementation of a prosody module in a TTS application.

Our experiments help advance research on the IS–prosody interface from an empirical perspective, and furthermore serve as a promising proof of concept that MTT's thematicity structure interpretation has a potential application for prediction of prosody in TTS applications, especially in the case of sentences of a certain syntactic complexity. Our experiments also show that mean acoustic parameters have a substantial potential for the prediction of hierarchical thematicity labels (even on an unbalanced corpus with low representativeness of some labels), in particular, for embedded thematicity spans. This suggests that the correlation between thematicity and prosody is bidirectional and thus the thematicity⟶prosody direction can be used to render the communicative intention of the speaker (e.g., a conversational agent) in terms of prosody, and the prosody⟶thematicity direction can be used to deduce the communicative intention of the speaker (e.g., a human user of the application). Further experiments on much larger corpora need to be carried out to further confirm this hypothesis.

Despite these advances, we must be aware that we are still a long way from achieving realistic prosody generation, since the highly complex issue of formal automated IS codification and analysis is still completely in its infancy. The presented study is furthermore also limited in itself with respect to (i) the corpus size and (ii) the text register, as it involves only read speech. But it has the advantage to be empirically grounded and to use a formal representation of thematicity that is independent of a question–answer setting. The fact that our corpus includes participants from different dialectal regions in the USA (as well as gender and age ranges) is admittedly a handicap to obtain a high level of coincidence in prosodic patterns. It is well-known that socio-cultural and dialectal variations affect prosody, and even though this is a really interesting field of research, it is out of our scope in this study to explore how those variables affect the IS–prosody correspondence.

Future work must address the compilation of corpora in other languages and registers following a semi-automatic methodology, as well as the annotation of other communicative dimensions of MTT's communicative structure such as, *foregroundedness*, *emphasis* and *focalization* (Mel'čuk, 2001) with the goal of the investigation of their joint correspondence to prosody. Only when we consider all dimensions together, will we get a more complete picture of the correspondence of the structural linguistic description with prosody and capture phenomena that in some other frameworks are considered as one – as, for example, *contrastive* vs. *non-contrastive topic* in Hedberg & Sosa (2008) (which in terms of Melčuk would be 'theme' + 'focalized' and 'theme' + 'non-focalized'). Another topic that we excluded from the present work and that is of utmost relevance in the context of prosody is the formal representation of pragmatic features and their relation to prosody. Finally, a topic that has been also clearly beyond the scope of our work so far is the determination of the IS in discourse Riester et al. (2018), which might complement Mel'čuk (2001)'s criteria for the determination of the thematic structure distribution.

Acknowledgements

This work has been partially funded by the European Commission in the context of its H2020 Programme under the contract numbers H2020-645012-RIA (KRISTINA) and H2020-870930-IA (WELCOME). The second author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE), grant RYC-2015-17239 (AEI/FSE, UE).

References

Ballesteros, M., Bohnet, B., Mille, S., & Wanner, L. (2015). Data-driven sentence generation with non-isomorphic trees. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL–HLT)* (pp. 387 - 397). Denver, Colorado: Association for Computational Linguistics.

Baumann, S. (2012). *The intonation of givenness: Evidence from German*. Tübingen: Max Niemeyer Verlag.

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

Black, A. W., & 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.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341 - 345.

Bohnet, B., Burga, A., & Wanner, L. (2013). Towards the annotation of Penn TreeBank with information structure. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 1250 - 1256). Nagoya, Japan: Association for Computational Linguistics.

Bouayad-Agha, N., Casamayor, G., Mille, S., & Wanner, L. (2012). Perspective-Oriented Generation of Football Match Summaries: Old Tasks, New Challenges. *ACM Transactions on Speech and Language Processing*, *9*, 1 - 31.

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

Büring, D. (2003). On D-trees, beans, and B-accents. *Linguistics and philosophy*, *26*, 511 - 545.

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

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

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and Topic* (pp. 25 - 55). New York: Academic Press.

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., & Johnson, M. (2000). BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43. URL: `https://www.cis.upenn.edu/~treebank/`.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse Production and Comprehension* (pp. 1 - 40). Norwood, New Jersey: Ablex Publishing Corporation volume 1 of *Discourse Processes: Advances in Research and Theory*.

Daneš, F. (1970). One instance of Prague School methodology: Functional analysis of utterance and text. In P. L. Garvin (Ed.), *Method and Theory in Linguistics* (pp. 132 - 146). Berlin, Germany: De Gruyter Mouton volume 40 of *Janua Linguarum. Series Maior*.

Domínguez, M., Burga, A., Farrús, M., & Wanner, L. (2018). Towards expressive prosody generation in tts for reading aloud applications. In *Proceedings of IberSPEECH 2018* (pp. 40 - 44). Barcelona, Spain.

Domínguez, M., Farrús, M., & Wanner, L. (2017). A thematicity-based prosody enrichment tool for CTS. In *Proceedings of Interspeech: show and tell demonstrations* (pp. 3421 - 3422). Stockholm, Sweden.

Domínguez, M., Latorre, I., Farrús, M., Codina, J., & Wanner, L. (2016). Praat on the web: An upgrade of praat for semi-automatic speech annotation. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 218 - 222). Osaka, Japan.

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

Grabe, E., Nolan, F., & Farrar, K. (1998). IViE – A comparative transcription system for intonational variation in English. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (pp. 1259 - 1262). Sydney, Australia.

Haji-Abdolhosseini, M. (2003). A constraint-based approach to information structure and prosody correspondence. In S. Müller (Ed.), *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar* (pp. 143 - 162). Michigan State University: CSLI Publications.

Hajičova, E. (1987). Focussing—A meeting point of Linguistics and Artificial Intelligence. In P. Jorrand, & V. Sgurev (Eds.), *Proceedings of the 2nd International Conference on Artificial Intelligence II: Methodology, Systems, Applications* (pp. 311 - 321). Varna, Bulgaria.

Hajičova, E., Partee, B., & Sgall, P. (1998). *Topic-Focus Articulation, Tripartite Structures, and Semantic Content* volume 71 of *Studies in Linguistics and Philosophy*. Dordrecht, Netherlands: Springer Netherlands.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*.

Halliday, M. (1967). Notes on Transitivity and Theme in English: Parts 1-3. *Journal of Linguistics*, *3*, 199 - 244.

Hedberg, N., & Sosa, J. (2008). The prosody of topic and focus in spontaneous English dialogue. In C. Lee, M. Gordon, & D. Büring (Eds.), *Topic and Focus. Studies in Linguistics and Philosophy*. Dordrecht, Netherlands: Springer volume 82.

Hirschberg, J. (2008). Pragmatics and intonation. In L. R. Horn, & G. Ward (Eds.), *The Handbook of Pragmatics* chapter 23. (pp. 515–537). John Wiley & Sons, Ltd.

Hirst, D., & Di-Cristo, A. (Eds.) (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge, United Kingdom: Cambridge University Press.

Izzad, R., Noraini, S., Norizah, A., & Nursuriati, J. (2016). Rule-based storytelling text-to-speech (TTS) synthesis. In *3rd International Conference on*

Mechanics and Mechatronics Research (ICMMR) (pp. 1 - 6). volume 77 of *MATEC Web Conferences*.

Kalbertodt, J., Primus, B., & Schumacher, P. B. (2015). Punctuation, prosody, and discourse: Afterthought vs. right dislocation. *Frontiers in Psychology*, *6*, 1 - 12.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*, 243 - 276.

Kruijff-Korbayová, I., Ericsson, S., Rodríguez, K. J., & 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)* (pp. 227 - 234). Budapest, Hungary.

Kügler, F., Smolibocki, B., & Stede, M. (2012). Evaluation of information structure in speech synthesis: The case of product recommender systems perception. In *ITG Symposium on Speech Communication* (pp. 26 - 29). Braunschweig, Germany: IEEE.

Ladd, R. (2008). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Levelt, W. (1993). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

Levitan, R., Beňuš, S., Gálvez, R. H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., & Hirschberg, J. (2016). Implementing acoustic-prosodic entrainment in a conversational avatar. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 1166 - 1170). San Francisco, USA.

López-Mencía, B., Díaz-Pardo, D., Hernández-Trapote, A., & Hernández-Gómez, L. A. (2013). Embodied conversational agents in interactive applications for children with special educational needs. In D. Griol Barres, Z. Callejas Carrión, & R. L.-C. Delgado (Eds.), *Technologies for Inclusive Education: Beyond Traditional Integration Approaches* (pp. 59 - 88). Hershey, USA: IGI Global.

Mathesius, V. (1929). Zur Satzperspektive im modernen Englisch. In *Archiv für das Studium der neueren Sprachen und Literaturen* (pp. 202 - 210). Erich Schmidt Verlag volume 155.

Mel'čuk, I. A. (2001). *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Amsterdam, Philadephia: Benjamins.

Meurers, D., Ziai, R., Ott, N., & 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 (pp. 1 - 9). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ortiz, A., del Puy Carretero, M., Oyarzun, D., Yanguas, J. J., Buiza, C., González, M. F., & Etxeberria, I. (2007). Elderly users in ambient intelligence: Does an avatar improve the interaction? In C. Stephanidis, & M. Pieper (Eds.), *Universal Access in Ambient Intelligence Environments: 9th ERCIM Workshop on User Interfaces for All* (pp. 99 - 114). Berlin, Heidelberg: Springer Berlin Heidelberg.

Pérez-Marín, D., & Pascual-Nieto, I. (2013). An exploratory study on how children interact with pedagogic conversational agents. *Behaviour & Information Technology*, *32*, 955–964.

Riester, A., Brunetti, L., & De Kuthy, K. (2018). Annotation guidelines for questions under discussion and information structure. In E. Adamou, K. Haude,

& M. Vanhove (Eds.), *Information Structure in Lesser-described Languages: Studies in Prosody and Syntax* (pp. 403 - 443). Benjamins.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, *1*, 75 - 116.

Schröder, M., & Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, *6*, 365 - 377.

Schwarzschild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent*. *Natural Language Semantics*, *7*, 141 - 177.

Selkirk, E. O. (1984). *Phonology and Syntax: The relation between sound and structure*. Cambridge, Massachussetts: The MIT Press.

Sgall, P., Hajičová, E., & Benešová, E. (1973). *Topic, focus and generative semantics*. Kronberg im Taunus, Germany: Scriptor.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. In *2nd International Conference on Spoken Language Processing (ICSLP 92)* (pp. 867 - 870). Banff, Canada.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, *31*, 649 - 689.

Syrdal, A. K., & Kim, Y.-J. (2008). Dialog speech acts and prosody : Considerations for TTS. In *Proceedings of Speech Prosody* (pp. 661 - 665). Campinas, Brazil.

Vallduví, E. (2016). Information structure. In M. Aloni, & P. Dekker (Eds.), *The Cambridge Handbook of Formal Semantics* Cambridge Handbooks in Language and Linguistics (pp. 728—-755). Cambridge University Press.

Vanrell, M., Mascaró, I., Torres-Tamarit, F., & Prieto, P. (2013). Intonation as an encoder of speaker certainty: Information and confirmation yes-no questions in Catalan. *Language and Speech*, *56*, 163 - 190.

Von Stechow, A. (1981). Topic, Focus and Local Relevance. In W. Klein, & W. Levelt (Eds.), *Crossing the Boundaries in Linguistics: Studies Presented to Manfred Bierwisch* (pp. 95 - 130). Dordrecht, Netherlands: Springer.

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., & 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.

Wargnier, P., Carletti, G., Laurent-Corniquet, Y., Benveniste, S., Jouvelot, P., & Rigaud, A. S. (2016). Field evaluation with cognitively-impaired older adults of attention management in the embodied conversational agent louise. In *2016 International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1 - 8). Orlando, FL, USA: IEEE.

Wolff, S., & Brechmann, A. (2015). Carrot and stick 2.0: The benefits of natural and motivational prosody in computer-assisted learning. *Computers in Human Behavior*, *43*, 76 - 84.