IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Towards expressive prosody generation in TTS for reading aloud applications

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

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Catalan Institue for Research and Advanced Studies (ICREA), Barcelona, Spain
monica.dominguez@upf.edu, alicia.burga@upf.edu, mireia.farrus@upf.edu, leo.wanner@upf.edu

Abstract

Conversational interfaces involving text-to-speech (TTS) applications have improved expressiveness and overall naturalness to a reasonable extent in the last decades. Conversational features, such as speech acts, affective states and information structure have been instrumental to derive more expressive prosodic contours. However, synthetic speech is still perceived as monotonous, when a text that lacks those conversational features is read aloud in the interface, i.e. it is fed directly to the TTS application. In this paper, we propose a methodology for pre-processing raw texts before they arrive to the TTS application. The aim is to analyze syntactic and information (or communicative) structure, and then use the high-level linguistic features derived from the analysis to generate more expressive prosody in the synthesized speech. The proposed methodology encompasses a pipeline of four modules: (1) a tokenizer, (2) a syntactic parser, (3) a communicative parser, and (3) an SSML prosody tag converter. The implementation has been tested in an experimental setting for German, using web-retrieved articles.

Index Terms: speech synthesis, thematicity, prosody, communicative structure, information structure.

1. Introduction

The advances in the development of Embodied Conversational Avatars (ECAs) in the last decades have allowed the exploration of increasingly complex communicative strategies in the field natural language interaction. Contextual information (such as profile information with respect to culture, age and even emotional state) is being used in the process of dialogue management, reasoning and expressive speech generation for the introduction of ECAs in social settings, for instance as social companions for children or elderly, or as health-care assistants [1, 2, 3]. In this context, the expressiveness of the generated speech plays an important role, which tends to be overlooked. Such expressiveness is mostly conveyed by means of prosody and usually involves an intertwined set of linguistic and paralinguistic functions that are difficult to grasp in a computational model.

Recent studies have demonstrated that high-level linguistic structures such as information (or communicative) structure, produced by the natural language generation (NLG) module in the course of sentence or text generation, are instrumental in the derivation of a more varied prosody in text-to-speech (TTS) applications [4]. When such features are not available, synthetic speech is perceived as monotonous, especially in long monologue discourse as observed in reading aloud exercises. In this case, raw text material (for instance, web-retrieved information) is passed as input to the TTS without any contextual or linguistic information.

In what follows, we propose a methodology for enriching web-retrieved texts with prosodic cues in the context of the *Knowledge-Based Information Agent with Social Competence and Human Interaction Capabilities* (KRISTINA [3]) in its role as reader. The cues are derived from the Information Structure, which is derived automatically for the discourse that is to be read – based on studies which show that when prosody reflects the Information Structure of a sentence, the overall comprehension of the message improves; see, e.g. [5] for German.

In the early 2000ies, there were some attempts to introduce some basic concepts of Information Structure in TTS applications, in particular, thematicity, understood as the partition of a sentence into *theme* (i.e., what the sentence in about) and *rheme* (i.e., what is said about the theme); see [6, 7] among others. However, a binary flat representation of thematicity of this kind has been proved to be insufficient to describe long complex sentences, whereas the hierarchical tripartite approach proposed in [8] within the Meaning–Text Theory yields a better correspondence to prosodic patterns as shown in [9, 10].

We introduce an experimental setup that targets reading aloud of a newspaper article in German. The setup is based on the sentence as the linguistic referent unit and on a formal representation of thematicity that follows the principles of the Meaning–Text Theory. The implementation of the proposed thematicity-based prosody enrichment encompasses four modules, one of which is off-the-shelf (i.e., the syntactic parser) and the rest have been developed by the authors: (1) a tokenizer, which prepares the input needed for the parser; (2) a communicative parser, which derives thematicity spans from morpho-syntactic features; and (3) an SSML prosody tag converter, which assigns a variety of SSML prosody tags based on the thematicity structure of each sentence.

The rest of the paper is structured as follows: Section 2 motivates this work, sets the context of the problem and refers to previous work in this area. In Section 3, we present the proposed methodology to automatically derive communicative structure from text and, thereupon, to generate prosodic contours using SSML tags. Section 4 introduces a sample implementation using a female German voice in MaryTTS. The output of this implementation is evaluated by means of a perception test in Section 5. Finally, conclusions are drawn in Section 6.

2. Motivation and Background

Different linguistic schools have long stated that Information Structure (IS), and, in particular, the dichotomy referred to as *theme–rheme* [11], *given–new* [12], or *topic–focus* [13] is related to intonation.[1] Moreover, prosody structure on the grounds of thematicity partitions plays a key role in the understanding of a message [14]. Empirical studies in different languages provide evidence that when thematicity and prosody are

[1]In our work, we use the first denotation, i.e., *theme–rheme* or *thematicity*.

10.21437/IberSPEECH.2018-9

appropriately correlated with each other, comprehension of the message is positively affected (cf., e.g., [5] for German and [15] for Catalan). Therefore, there is reason to assume that a conversational application considering the notions of content packaging by means of the relation between thematicity and prosody will benefit from the same advantages as in natural conversation environments. Most of all, conversational avatars in applications for children in educational settings [16], applications for those with special needs [1] as well as for elderly [2] and, in particular, for those with cognitive impairments [17], would greatly benefit from such a communicatively-oriented improvement.

State-of-the-art conversational applications, in particular TTS systems, do not yet include communicative information. The task is not trivial. It involves, in the first place, a communicative theoretical model, automatic tools to parse the Information Structure of a text, and, last but not least, a generative model of the related prosodic contour. Some preliminary attempts to include thematicity in TTS applications were made in the past. Consider, for instance, Steedman's work [6] on the correlation of theme and rheme to rising and falling intonation patterns, which was tested in the Festival speech synthesizer [18], and the creation of dedicated tags in MaryTTS [19] for the notions of *givenness* and *contrast* [20]. However, these attempts have a major shortcoming in that they use a flat binary thematicity structure, which does not suffice to describe the complexity of content packaging, especially in relation to prosody. Our previous studies (see, e.g. [21]) suggest that hierarchical thematicity based on propositions as described by Mel'čuk [8] constructs a more versatile scaffolding for communicative modeling of computer interaction with humans. As already mentioned above, Mel'čuk's methodological approach has also demonstrated to be instrumental in natural language generation applications [22, 23].

In contrast to IS models that propose a partition of sentences into a theme and a rheme, Mel'čuk [8] argues in the context of the Meaning–Text Theory for a tripartite hierarchical division ('theme', 'rheme', and 'specifier' – the element which sets the utterance's context) within propositions that further permits embeddedness of communicative spans; consider (1) for illustration of hierarchical thematicity (annotated following the guidelines established in [24]) of the sentence *Ever since, the remaining members have been desperate for the United States to rejoin this dreadful group*. A total of five partitions are identified, including three spans at level 1 (L1), a specifier (SP1), theme (T1) and rheme (R1), and two embedded spans at level 2 (L2)[2] in the rheme, a theme (T1(R1)) and a rheme (R1(R1)).[3]

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

A hierarchical thematicity structure of this kind has been shown to correlate better with ToBI labels than binary flat thematicity [9]. Such a correlation still does not solve the problem of a one–to-one mapping between a specific intonation label (e.g., H*) to a static acoustic parameter (e.g., an increase of 50% in fundamental frequency). A more varied range of prosodic cues based on the analysis of the available corpus of read speech annotated with hierarchical thematicity (in the NLG module) and prosody has been proved to yield an improvement in the perception of expressiveness of the synthesized speech [4]. However, there is still no application that can provide an automatic derivation of thematicity-based prosody cues for raw texts that arrive to the TTS application, such as the implementation proposed in this paper.

[2]Levels are connected to the concept of embeddedness of spans: for instance, a main theme (T1 at L1) may be subdivided into further thematicity spans, which will belong to L2 thematicity.

[3]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).

3. Methodology

This paper proposes an approach that tests the formal representation of information (or communicative) structure proposed by Mel'čuk [8] and its correspondence to prosody in the context of a concept-to-speech (CTS) application, where text coming from a web-retrieved service is input to the TTS engine.

3.1. Objectives

Our work envisages the study of the IS–prosody interface from a methodological perspective based on a speech synthesis implementation setup. The proposed methodology has the following underlying goals:

- to provide automatic tools to investigate the effect of thematicity–prosody correspondence in human-machine interaction contexts;
- to explore the advantages and limitations of a thematicity-based prosody enrichment in speech synthesis;
- to provide a preliminary scaffolding to incrementally add other communicative dimensions, registers and languages;

Such a methodology addresses two main research issues in this field: (i) the lack of implementation settings of the IS–prosody correspondence and (ii) testing of the integration of the IS–prosody interface in computational settings.

3.2. Pipeline

The proposed pipeline sketched in Figure 1 includes four modules:

1. **Tokenizer**: it splits the text into sentences and words. Punctuation marks are also tokenized as required to serve as input for the syntactic parser.
2. **Syntactic parser**: The parser by Bohnet [25, 26] is used. This parser is trained on the TIGER Penn Treebank [27] and outputs a fourteen-columned CONLL file.
3. **Communicative parser**. This rule-based system derives thematicity labels from syntactic structure. It outputs a CONLL file with an added column for communicative structure (i.e., the output CONLL has fifteen columns). For now, it only derives hierarchical thematicity labels.
4. **SSML prosody converter**. It converts the thematicity spans derived by the communicative parser to SSML spans and assigns a variety of prosody tags to each span. This module is based on the tool presented in [28].

The use of the *Speech Synthesis Markup Language* (SSML) [29] convention for prosody enrichment, as proposed in [28], facilitates the integration of the methodology proposed in this paper within the context of TTS applications.

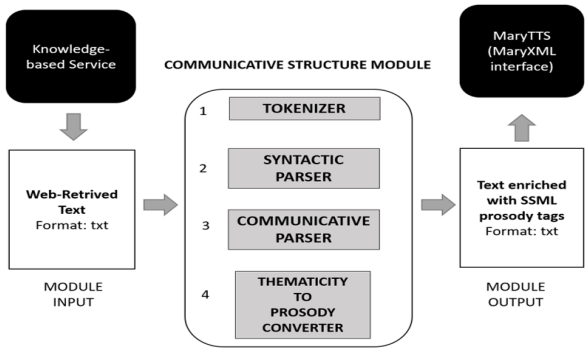

Figure 1: *Communicative generation pipeline.*

1	Der	_	der	_	ART	_	Nom|Sg|Masc	-1	2	_	NK	_	_	[
2	Begriff	_	begriff	_	NN	_	Nom|Sg|Masc	-1	5	_	SB	_	_	_
3	der	_	der	_	ART	_	Gen|Sg|Fem	-1	4	_	NK	_	_	_
4	Schlafhygiene	_	schlafhygiene	_	NN	_	Gen|Sg|Fem	-1	2	_	AG	_	_	]T1
5	bezeichnet	_	bezeichnen	_	VVFIN	_	3|Sg|Pres|Ind	-1	0	_	ROOT	_	_	[
6	Verhaltensweisen	_	verhaltensweise	_	NN	_	Dat|Pl|Masc	-1	5	_	OA	_	_	_
7	,	_	_	_	$,	_	_	-1	6	_	PUNC	_	_	_
8	die	_	der	_	PRELS	_	Nom|Sg|Fem	-1	12	_	SB	_	_	[
9	einen	_	ein	_	ART	_	Acc|Sg|Masc	-1	11	_	NK	_	_	_
10	gesunden	_	gesund	_	ADJA	_	Pos|Acc|Sg|Masc	-1	11	_	NK	_	_	_
11	Schlaf	_	schlaf	_	NN	_	Acc|Sg|Masc	-1	12	_	OA	_	_	_
12	fördern	_	fördern	_	VVFIN	_	Inf	-1	6	_	RC	_	_	]P2]R1
13	.	_	_	_	$.	_	_	-1	5	_	PUNC	_	_	_

Figure 2: *Example of the output of the communicative parser in CONLL format annotated with thematicity.*

3.3. The Communicative Parser

The main contribution of this paper is a rule-based communicative parser for texts in German. In what follows, we sketch the core functions and algorithm of the parser.[4]

The parser is implemented as a python script that requires a CONLL file with the part-of-speech (POS) and dependency syntax analysis per token in each sentence. Clauses are the main syntactic cue to detect propositions. Thus, The main algorithm loops over POS and dependency relations columns to identify complex and coordinated clauses in the first place and label propositions. Then, thematicity is labeled focusing on the detection of specifiers and themes, which usually have as syntactic correlates frontal modifiers and subjects respectively.

Several functions have been scripted for finding propositions, thematicity and annotate them following the guidelines established in [24]. Those guidelines establish the convention of using square brackets ("[" and "]") to establish the beginning and end of a thematicity span and keys ("{" and "}") to signal beginning and end of a proposition. The resulting output is a CONLL file that has one column at the end with the annotation of thematicity (cf. Figure 2).

4. Experimental Setup

A working corpus has been created from web-retrieved text in German on advice for sleeping routines and local news. The corpus contains eight texts with a total of 1,418 words.In what follows, we present the experimental setup with respect to the prosody enrichment procedure and the IS–prosody correspondence.

The open source software MaryTTS[5] [19] was used for the implementation. The default synthesized speech output has been enriched using MaryXML prosody specifications[6], which follow the SSML recommendation[7].

The SSML *prosody tags* allow 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., '+50Hz' for increasing a specific amount of hertz (Hz) in F0) and relative values can be used to applied the modification. An example of a SSML prosody tag for modification of two prosodic elements is presented below:

Example (1)

<prosody rate="-10%" pitch="+20%">*text to be modified* </prosody>

Moreover, the SSML *boundary tag* that controls the introduction of pauses at a specific location was also used after each thematicity span. The duration of the break is specified in milliseconds (ms). An example of SSML boundary tag is introduced below:

Example (2)

Text before the break <boundary duration="100"/>*text after the break.*

The correspondence between thematicity and prosody is presented as variations from referent prosody tag values involving fundamental frequency (F0) and speech rate (SR) over thematicity spans (cf. Table 1). We propose testing a varied range of values generated automatically, against a manual implementation following the findings in [10, 4], where a variety of prosodic cues for each thematicity span is presented based on corpus analysis.

	F0	SR
T1	+15%	-15%
R1	+10%	+10%
SP1	+20%	-10%
P	+15%	-10%

Table 1: *Referent prosody tag values for L1 thematicity.*

Table 1 shows the referent modification for theme (T1), rheme (R1) and specifier (SP1) spans within L1 thematicity. Propositions are defined as clauses that contain a finite verb and they are the referent units for thematicity segmentation. They can include L1 and L2 spans and embrace under one communicative label different types of syntactic relationships, for example coordination, juxtaposition and subordination. The referent values assigned to each span are chosen randomly within a range of plus minus 5 points in each new sentence. Thus, even though the annotation of thematicity in this experiment is restricted to the sentence domain, an automatic variation is envisaged to generate a different range of prosodic parameters across sentences.

[4]The code of both the communicative parser and the thematicity to SSML module is available in the following repository under a GNU v.3 licence: https://github.com/TalnUPF/KRISTthem2prosModule

[5]Available at http://mary.dfki.de/

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

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

5. Evaluation

For the evaluation of automatic assignment of thematicity-based prosody, a selection of newspaper articles in German has been done. From those articles, a selection of sentences with different communicative structures has been made for the perception test, as detailed below.

For the evaluation of the thematicity-based prosody enrichment module, expressiveness was assessed by means of a perception test using: (1) a Mean Opinion Score (MOS) with a 5-point Likert scale: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent; and a pairwise comparison. Seventeen participants took part in the evaluation, all of them either native speakers of German or proficient speakers. The test was conducted fully in German and participants were informed that our goal is to investigate if synthesized speech was perceived as better expressing the communicative content of the sentence taking into account prosodic variability. Six sentences were included in the perception test representative of different complexity in syntax and communicative structure:

S1 *Warme Fuß und Vollbäder direkt vor dem Schlafengehen fördern den Nachtschlaf.*

S2 *Der Begriff der Schlafhygiene bezeichnet Verhaltensweisen, die einen gesunden Schlaf fördern.*

S3 *Dafür sorgen, dass das Schlafzimmer ruhig und dunkel ist und eine angenehme Temperatur hat.*

S4 *Landrat Thomas Reumann schlägt vor, den Finanzierungsantrag zu stellen, will aber erst im Haushalt 2018 Gelder einstellen.*

S5 *Das funktioniert nur, wenn alle mitmachen.*

S6 *Im übrigen betonte er, dass der Landkreis nicht allein sei, sondern Städte und Gemeinden als Partner habe, die den Beschluss mittragen müssten.*

Three samples of each sentences were included in the MOS test: (1) the default TTS output (DEF); (2) automatic thematicity-based modifications (AUT) and (3) manual thematicity-based prosody modifications (MAN). The pairwise comparison included the default TTS output versus the automatic thematicity-based prosody modification. A total of fifty-one answers are considered in the evaluation. Table 2 shows results of the MOS test. In all cases, the best scoring sample is the thematicity-based prosody modification (either manual or automatic). This supports the initial hypothesis that thematicity-based prosody modifications are perceived as more expressive. In sentences 2, 3, 5 and 6 the best scoring option is the automatic version, whereas sentences 1 and 4 score best for the manual version of the modification. These results are in line with the pairwise comparison shown in Table 3, where all choices go for the thematicity-based modification except for sentence 4.

	S1	S2	S3	S4	S5	S6	Average
DEF	2.65	3.35	3.12	2.71	3.18	3.06	3.01
AUT	3.00	3.53	3.41	2.65	3.53	3.71	3.30
MAN	3.12	2.88	3.06	3.00	2.71	2.88	2.94

Table 2: *Results from the MOS test for the thematicity-based prosody enrichment module*

A t-test shows that overall (average) results for the automatic prosody modifications (AUT = 3.30) achieve statistical significance at $p < 0.05$ compared to the default score (DEF = 3.01).

	S1	S2	S3	S4	S5	S6	Average
DEF	47%	29%	29%	71%	41%	47%	44%
AUT	53%	71%	71%	29%	59%	53%	56%

Table 3: *Results from the pairwise comparison for the default and automatic modification*

6. Conclusions

Given the relevant role of the Information Structure–prosody interface in human communication, it seems reasonable that next generation conversational agents 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 capabilities of prosody enrichment standards to achieve variability in implementation settings.

The present study provides a methodology for a more versatile integration of the IS–prosody interface in TTS for reading aloud applications. Such a methodology contributes in several aspects to the state of the art: (i) a formal description of hierarchical thematicity is used; (ii) a communicative parser that derives thematicity labels is introduced; and (iii) the prosodic cues are automatically derived and tested in a TTS application. All in all, this study pivots the transition from theoretical work on the IS–prosody interface to the integration of thematicity-based prosody enrichment to achieve more expressive synthesized speech.

A 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 remain to be looked into. Future work is, furthermore, aimed at exploring other dimensions of communicative structure like emphasis and foregroundedness within the framework that has been proposed in this paper.

7. Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Unions Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. It has been also partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). The third author is partially funded by the *Ramón y Cajal* program.

8. References

[1] B. L. Mencía, D. D. Pardo, A. H. Trapote, and L. A. H. Gómez, "Embodied Conversational Agents in Interactive Applications for Children with Special Educational Needs," in *Technologies for Inclusive Education: Beyond Traditional Integration Approaches*, D. Griol Barres, Z. Callejas Carrión, and R. L.-C. Delgado, Eds. Hershey, USA: IGI Global, 2013, pp. 59–88.

[2] A. Ortiz, M. del Puy Carretero, D. Oyarzun, J. J. Yanguas, C. Buiza, M. F. Gonzalez, and I. Etxeberria, *Elderly Users in Ambient Intelligence: Does an Avatar Improve the Interaction?* Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 99–114.

[3] L. Wanner, E. André, J. Blat, S. Dasiopoulou, M. Farrús, T. Fraga, E. Kamateri, F. Lingenfelser, G. Llorach, O. Martínez, G. Meditskos, S. Mille, W. Minker, L. Pragst, D. Schiller, A. Stam, L. Stellingwerff, F. Sukno, B. Vieru, and S. Vrochidis, "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, 2017.

[4] M. Domínguez, M. Farrús, and L. Wanner, "Thematicity-based Prosody Enrichment for Text-to-Speech Applications," in *Proceedings of the 9th International Conference on Speech Prosody 2018 (SP2018)*, Poznań, Poland, 2018.

[5] D. Meurers, R. Ziai, N. Ott, and J. Kopp, "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*, ser. TIWTE '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1–9.

[6] M. Steedman, "Information structure and the syntax-phonology interface," *Linguistic inquiry*, vol. 31, no. 4, pp. 649–689, Fall 2000.

[7] M. Haji-Abdolhosseini and S. Müller, "Constraint-Based Approach to Information Structure and Prosody Correspondence," in *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*. CSLI Publications, 2003, pp. 143–162.

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

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

[10] M. Domínguez, M. Farrús, and L. Wanner, "Compilation of corpora to study the information structureprosody interface," in *11th edition of the Language Resources and Evaluation Conference (LREC2018)*, Mijazaki, Japan, 2018.

[11] M. Halliday, "Notes on Transitivity and Theme in English, Parts 1-3," *Journal of Linguistics*, vol. 3, no. 1, pp. 37–81, 1967.

[12] R. Schwarzschild, "Givenness, avoidf and other constraints on the placement of accent," *Natural Language Semantics*, vol. 7, no. 1, pp. 141–177, 1999.

[13] E. Hajičová, B. Partee, and P. Sgall, *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht, 1998.

[14] H. H. Clark and S. E. Haviland, "Comprehension and the given-new contract," *Discourse production and comprehension. Discourse processes: Advances in research and theory*, vol. 1, pp. 1–40, 1977.

[15] M. Vanrell, I. Mascaró, F. Torres-Tamarit, and P. Prieto, "Intonation as an Encoder of Speaker Certainty: Information and Confirmation Yes-No Questions in Catalan," *Language and Speech*, vol. 56, no. 2, pp. 163–190, 2013.

[16] D. Prez-Marn and I. Pascual-Nieto, "An exploratory study on how children interact with pedagogic conversational agents," *Behaviour & Information Technology*, vol. 32, no. 9, pp. 955–964, 2013.

[17] P. Wargnier, G. Carletti, Y. Laurent-Corniquet, S. Benveniste, P. Jouvelot, and A. S. Rigaud, "Field evaluation with cognitively-impaired older adults of attention management in the Embodied Conversational Agent Louise," in *IEEE International Conference on Serious Games and Applications for Health, SeGAH 2016*, Orlando, Florida, USA, 2016, pp. 1–8.

[18] A. W. Black and P. A. Taylor, "The Festival Speech Synthesis System: System documentation," Human Communciation Research Centre, University of Edinburgh, Scotland, UK, Tech. Rep. HCRC/TR-83, 1997, avaliable at http://www.cstr.ed.ac.uk/projects/festival.html.

[19] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003. [Online]. Available: http://mary.dfki.de

[20] F. Kügler, B. Smolibocki, and M. Stede, "Evaluation of information structure in speech synthesis : The case of product recommender systems perception," in *ITG Conference on Speech Communication, IEEE*, 2012, pp. 26–29.

[21] M. Domínguez, M. Farrús, and L. Wanner, "Combining acoustic and linguistic features in phrase-oriented prosody prediction," in *Proceedings of the 8th International Conference on Speech Prosody*, Boston, USA, 2016, pp. 796–800.

[22] L. Wanner, B. Bohnet, and M. Giereth, "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*, 2003, pp. 100–104.

[23] M. Ballesteros, B. Bohnet, S. Mille, and L. Wanner, "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)*, 2015.

[24] B. Bohnet, A. Burga, and L. Wanner, "Towards the annotation of penn treebank with information structure," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013, pp. 1250–1256.

[25] B. Bohnet and J. Nivre, "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, Jeju Island, Korea, 2012, pp. 1455–1465.

[26] ——, "he Best of BothWorlds – A Graph-based Completion Model for Transition-based Parsers," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France, 2012, pp. 77–87.

[27] S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit, "TIGER: Linguistic Interpretation of a German Corpus," *Journal of Language and Computation*, no. 2, pp. 597–620, 2004.

[28] M. Domínguez, M. Farrús, and L. Wanner, "A thematicity-based prosody enrichment tool for cts," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, Stockholm, Sweden, 2017, pp. 3421–2.

[29] P. Taylor and A. Isard, "SSML: A Speech Synthesis Markup Language," *Speech Communication*, vol. 21, no. 1-2, pp. 123–133, February 1997.