Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 95-98 recibido 12-11-2018 revisado 07-01-2019 aceptado 05-02-2019

# La interfaz estructura informativa-prosodia: el rol de la tematicidad jerárquica basado en un modelo empírico

## *The Information Structure–Prosody Interface: On the Role of Hierarchical Thematicity in an Empirically-grounded Model*

**Mónica Domínguez Bajo**
Universitat Pompeu Fabra
C. Roc Boronat, 138
08018 Barcelona
monica.dominguez@upf.edu

**Resumen:** Tesis doctoral en tecnologías de la información y las comunicaciones realizada por Mónica Domínguez Bajo en la Universidad Pompeu Fabra bajo la dirección del Dr. Leo Wanner y la Dra. Mireia Farrús Cabecerán. El acto de defensa tuvo lugar el viernes 17 de noviembre de 2017 ante el tribunal formado por los doctores Bernd Möbius (Universidad de Saarland), Pilar Prieto Vives (Universidad Pompeu Fabra) y Catherine Lai (Universidad de Edimburgo). Obtuvo la calificación de Sobresaliente.
**Palabras clave:** Estructura informativa, estructura comunicativa, tematicidad, tema, rema, prosodia, parámetros acústicos, síntesis de voz, texto a habla, concepto a habla, etiquetado automático de prosodia

**Abstract:** PhD thesis in communication and information tecnologies written by Mónica Domínguez Bajo at the University Pompeu Fabra under the supervision of Dr. Leo Wanner and la Dra. Mireia Farrús Cabecerán. The author was examined on Friday, 17th November 2017 by a committee formed by the doctors Bernd Möbius (University of Saarland), Pilar Prieto Vives (University Pompeu Fabra) y Catherine Lai (University of Edinburgh). It obtained the grade of Excellent.
**Keywords:** Information structure, communicative structure, thematicity, theme, rheme, prosody, acoustic parameters, speech synthesis, text-to-speech, concept-to-speech, automatic prosody labeling

### *1 Introducción*

Las tecnologías del habla han pasado en poco tiempo de desarrollar tareas de lectura simples, como el sistema MITalk, a mantener conversaciones con interlocutores humanos, como es el caso de aplicaciones en el ámbito de la salud[1]. A pesar de que estas interacciones son relativamente sencillas, los asistentes virtuales están consiguiendo un impacto considerable en nuestra sociedad. No obstante, en lo que se conoce como tecnologías de 'texto a habla' (TTS, por sus siglas en inglés) aún no se han llegado a integrar aspectos comunicativos que doten a la generación de habla sintética de la versatilidad que existe en el lenguaje natural.

La expresividad de las voces sintéticas a través de la modelización de prosodia en los TTS tiene en cuenta, hasta cierto punto, algunos rasgos y funciones lingüísticas, sin embargo, aún no se ha llegado a alcanzar la riqueza que la prosodia tiene en el lenguaje humano. Por este motivo las voces sintéticas se siguen percibiendo como monótonas, sobre todo en discurso monologado con frases largas y complejas. Especialmente en el contexto de las tecnologías conversacionales, se espera que los agentes virtuales sean capaces tanto de expresarse de una manera apropiada al contexto como de mantener el interés del interlocutor. Esto, en la actualidad, supone un reto que requiere de una actualización en la agenda de investigación para incluir aspectos comunicativos que no se están teniendo en cuenta en tecnologías del habla.

[1]Entre otros, el avatar conversacional KRISTINA: http://kristina-project.eu/en/

ISSN 1135-5948. DOI 10.26342/2019-62-11

Existe una amplia bibliografía en lingüística teórica que viene enfatizando: (i) que la prosodia desempeña un papel clave a la hora de expresar la intención comunicativa del hablante; (ii) que dicha intención comunicativa se articula en términos de la estructura informativa y (iii) que la estructura informativa se puede generar mediante un procedimiento computacional de organización del contenido semántico y sintáctico. A principios del siglo XXI se atisbaban modestos intentos de aplicar este conocimiento lingüístico en la práctica implementando conceptos básicos de la estructura informativa, principalmente, la segmentación en términos de tematicidad. La tematicidad describe cómo se empaqueta el contenido en función de "lo que se está hablando", el *tema*, y "lo que se dice al respecto", el *rema*. Los intentos de implementación de la tematicidad en TTS se basaron en una correspondencia básica entre tema-rema con patrones de entonación ascendentes-descendentes. Sin embargo, se estaban subestimando dos aspectos esenciales desde el punto de vista computacional: la asignación de la tematicidad dado un texto cualquiera y la generación de un abanico de rasgos prosódicos suficiente para aportar la variabilidad y expresividad necesarias.

El objetivo principal de esta tesis es demostrar empíricamente la viabilidad de una generación de prosodia en habla sintética que tenga en cuenta la intención comunicativa del mensaje. Con este fin, se propone una metodología para avanzar en el estudio de la interfaz estructura informativa-prosodia desde un punto de vista empírico para su aplicación en implementaciones de generación de habla sintética a partir de texto.

## *2 Organización de la tesis*

La tesis se estructura en siete capítulos incluyendo introducción y conclusiones. En esta sección, se presenta un breve resumen de los capítulos centrales, es decir, del capítulo 2 al 6.

El capítulo 2 incluye una descripción de los conceptos fundamentales de las distintas áreas que abarca la tesis, en concreto:

- Estructura informativa o comunicativa: se explica la teoría de Igor Mel'čuk, que introduce una representación formal para aplicaciones computacionales de generación en el ámbito del procesamiento del lenguaje natural. De este modo, se enmarca el objeto de estudio de la presente tesis: la tematicidad jerárquica, tal y como la define Mel'čuk. Se detalla cómo esta teoría identifica tres segmentos sobre proposiciones: tema, rema y especificador. Dichos segmentos son recursivos y por lo tanto pueden definirse a distintos niveles de tematicidad.
- Prosodia: se introduce la convención de etiquetado ToBI[2], así como los parámetros acústicos que se usan en los experimentos. También se explican los procedimientos y convenciones que se usan en los sintetizadores del estado del arte para generar y modificar la prosodia.

Se describen las estrategias de los TTS del estado del arte para derivar prosodia:

- Sistemas de reglas: emplean características lingüísticas de bajo nivel, por ejemplo, posición de las palabras, tipología (si es una palabra funcional o de contenido) y puntuación.
- Árboles de decisión: parten de información morfo-sintáctica, por ejemplo, si la palabra es un sustantivo o un verbo y qué palabras dependen de ella.
- Superposición de parámetros prosódicos: se emplean etiquetas usando un lenguaje de marcado para síntesis de habla, entre los más populares está la convención Speech Synthesis Markup Language (conocida por sus siglas en inglés, SSML)[3]. La manipulación de la prosodia basada en una simplificación de las etiquetas ToBI también se utiliza en algunos sintetizadores.

El capítulo 3 resume los estudios relacionados con la interfaz estructura informativa-prosodia desde el punto de vista de la lingüística computacional, aunque también se hace referencia a una selección de estudios teóricos. Por otro lado, se mencionan las herramientas informáticas existentes para anotar y analizar prosodia, sus ventajas y limitaciones.

En el capítulo 4, se explica la metodología propuesta, el corpus de trabajo y los procedimientos para anotar prosodia empleados

[2]Siglas en inglés correspondientes a *Tones and Breaks Indices*.

[3]https://www.w3.org/TR/speech-synthesis11/

mediante la convención ToBI y extrayendo parámetros acústicos automáticamente.

Los capítulos 5 y 6 detallan los experimentos realizados en el marco de estudio de la tesis. El capítulo 5 presenta los experimentos de anotación de prosodia así como el sistema desarrollado para realizarlo automáticamente. El capítulo 6 incluye los experimentos de análisis de la correspondencia estructura informativa y prosodia a través de pruebas estadísticas y aprendizaje automático. Así mismo, se presenta la implementación en el entorno de TTS para testear las conclusiones extraídas del análisis del corpus de estudio.

## *3 Contribución de la tesis*

La tesis contribuye al avance del estado del arte en la integración de la interfaz estructura informativa-prosodia en tecnologías del habla desde dos ámbitos: teórico y técnico.

### 3.1 Contribución teórica

Los experimentos de análisis basado en un corpus de habla leída en inglés americano confirman que existe una correspondencia entre tematicidad jerárquica y prosodia que se puede modelizar para enriquecer la generación de prosodia.

En primer lugar se realiza un análisis estadístico de los parámetros prosódicos relacionados con los elementos que definen la prosodia, a saber, las pausas o fraseología, frecuencia fundamental (F0), la intensidad y la velocidad de habla. A continuación, se justifica por qué la tematicidad jerárquica propuesta por Mel'čuk es más adecuada que la segmentación binaria que se venía empleando en aplicaciones computacionales. Se realizan experimentos de clasificación para comprobar el potencial de predicción de prosodia a partir de tematicidad y viceversa mediante algoritmos de aprendizaje automático usando dos representaciones de prosodia: con etiquetas ToBI y con parámetros acústicos.

Finalmente, los experimentos en habla sintética muestran que los sintetizadores del estado del arte, tanto comerciales como de código libre, no tienen en cuenta la estructura comunicativa. Se muestra que el etiquetado ToBI, que es la convención más usada en estudios de prosodia, es muy limitado para su uso en enriquecimiento de prosodia en sintetizadores debido a su poca flexibilidad. Por este motivo, se propone el uso de la convención SSML, que permite una flexibilidad mayor no solo en modificaciones de F0, sino también en intensidad, pausas y velocidad de habla.

### 3.2 Contribución técnica

En esta tesis se han desarrollado dos aplicaciones de código abierto, que se detallan a continuación.

#### 3.2.1 Etiquetador automático de prosodia

El etiquetador automático de prosodia, está disponible en un servicio web, Praat on the Web[4]. El código es abierto y se distribuye bajo Licencia GNU v3[5]. La herramienta es modular y utiliza una extensión de Praat para el etiquetado automático de muestras de habla con o sin alineación por palabras. Se trata de un sistema basado en reglas que anota prominencia y fraseología prosódicas teniendo en cuenta parámetros acústicos. La evaluación de esta herramienta se realiza con muestras de habla leídas y espontáneas tanto en español como en inglés.

#### 3.2.2 Módulo de enriquecimiento prosódico

El módulo de enriquecimiento prosódico basado en la tematicidad jerárquica se puede usar con aplicaciones TTS que interpreten etiquetas SSML. Esta herramienta permite testear los resultados del análisis empírico del corpus sobre la correspondencia tematicidad jerárquica-prosodia en el entorno de habla sintética. Por lo tanto, se promueve un estudio aplicado de la interfaz estructura informativa-prosodia, lo cual supone un paso de la teoría lingüística en este área a la práctica en el contexto de tecnologías del habla, que hasta ahora no se había realizado.

## *Agradecimientos*

La autora ha recibido financiación durante la tesis de las siguientes entidades: la Universidad Pompeu Fabra mediante una beca predoctoral del departamento de tecnologías de la información y las comunicaciones, el Ministerio de Economía y Competitividad mediante el programa de excelencia María de Maeztu (MDM-2015-0502) y la Comisión Europea a través del proyecto KRISTINA (H2020-RIA-645012).

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

[5] https://github.com/monikaUPF

## *Bibliografía*


Domínguez, M., A. Burga, M. Farrús, y L. Wanner. 2018. On the Role of Communicative Structure in Read Aloud Applications for the Elderly. En *Proceedings of the Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications, AAMAS'18*, Stockholm, Sweden.

Domínguez, M., M. Farrús, A. Burga, y L. Wanner. 2014. The Information Structure–Prosody Language Interface Revisited. En *Proceedings of the 7th International Conference on Speech Prosody*, páginas 539–543, Dublin, Ireland.

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

Domínguez, M., M. Farrús, y L. Wanner. 2016a. An Automatic Prosody Tagger for Spontaneous Speech. En *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, páginas 377–387, Osaka, Japan.

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

Domínguez, M., M. Farrús, y L. Wanner. 2017. A thematicity-based prosody enrichment tool for cts. En *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, páginas 3421–2, Stockholm, Sweden.

Domínguez, M., M. Farrús, y L. Wanner. 2018a. Compilation of corpora to study the information structure–prosody interface. En *11th edition of the Language Resources and Evaluation Conference (LREC2018)*, páginas 4030–4035, Mijazaki, Japan.

Domínguez, M., M. Farrús, y L. Wanner. 2018b. Thematicity-based Prosody Enrichment for Text-to-Speech Applications. En *Proceedings of the 9th International Conference on Speech Prosody 2018 (SP2018)*, páginas 612–616, Poznań, Poland.

Domínguez, M., M. Farrús, y L. Wanner. 2018c. Towards Expressive Prosody Generation in TTS for Reading Aloud Applications. En *Proceedings of IberSpeech 2018: International Speech Communication Association*, páginas 40–44, Barcelona, Spain.

Domínguez, M., I. Latorre, M. Farrús, J. Codina, y L. Wanner. 2016. Praat on the Web: An Upgrade of Praat for Semi-Automatic Speech Annotation. En *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, páginas 218–222, Osaka, Japan.

Domínguez, M., M. Farrús, A. Burga, y L. Wanner. 2014. Towards Automatic Extraction of Prosodic Patterns for Speech Synthesis. En *Proceedings of the 7th International Conference on Speech Prosody*, páginas 1105–1109, Dublin, Ireland.