DECEMBER | 13 - 14
Universitat de València

LingCor 2018

1st International Workshop on Spoken Corpus Linguistics

INVITED SPEAKERS

Anke Lüdeling
(Humboldt-Universität zu Berlin)

Antonio Briz Gómez
(Universitat de València)

Martin Hilpert
(Université de Neuchâtel)

Xosé Luís Regueira Fernández
(Universidade de Santiago de Compostela)

ORGANISATION

The **2018 LingCor Workshop** is organised within the framework of the research project "Elaboració d'un corpus oral dialectal del valencià col·loquial (CorDiVal)" (ref. GV/2017/094), funded by the Valencian Government.

SPONSOR

GENERALITAT VALENCIANA
Conselleria d'Educació, Investigació, Cultura i Esport

COLLABORATORS

Facultat de Filologia, Traducció i Comunicació

DEPARTAMENT FILOLOGIA CATALANA UNIVERSITAT D'ALACANT

GEVaD

lsi Departament de Llenguatges i Sistemes Informàtics

PROGRAMA

13 de desembre

9.00 a 9.15	Registre
9.15 a 9.30	Presentació
9.30 a 10.30	**Ponència. Anke Lüdeling** (Institut für Deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin) *Optionality and variation in spoken learner language*
10.30 a 11.00	Vicent Beltran (Universitat d'Alacant), Miquel Esplà (Universitat d'Alacant), Maribel Guardiola (Universitat d'Alacant), Sandra Montserrat (Universitat d'Alacant), Carles Segura (Universitat d'Alacant) & Andreu Sentí (Universitat de València) *A corpus of informal speech for Valencian Catalan: Parlars*
11.00 a 11.30	Descans
11.30 a 12.00	Anne Breitbarth (Ghent University), Anne-Sophie Ghyselen (Ghent University), Melissa Farasyn (Ghent University) & Jacques Van Keymeulen (Ghent University) *A parsed corpus of Southern Dutch dialects*
12.00 a 12.30	Giulia Bossaglia (Universidade Federal de Minas Gerais) & Lucia De Almeida Ferrari (Universidade Federal de Minas Gerais) *Methods, resources and some results of the C-ORAL-BRASIL project for Brazilian Portuguese spoken corpora*
12.30 a 13.00	Victoria Vázquez Rozas (Universidade de Santiago de Compostela) & Mario Barcala (NLPgo) *The ESLORA Corpus of Spoken Spanish: Design, Compilation and Search Engine*
13.00 a 13.30	Miriam Bouzouita (Ghent University) *In Search of New Transcription & Annotation Methods: The Collaborative Game-Based Approach*
13.30 a 15.00	Dinar
15.00 a 15.30	Adrián Cabedo Nebot (Universitat de València) *Problemas en el diseño de un corpus oral: el corpus Ameresco*
15.30 a 16:00	Marta Albelda Marco (Universitat de València) & Maria Estellés Arguedas (Universitat de València) *Problemas y soluciones en el diseño y construcción del corpus Ameresco*
16.00 a 16.30	Descans
16.30 a 17.30	**Ponència. Antonio Briz** (Universitat de València) *Los corpus de conversaciones coloquiales. La elaboración del corpus AMERESCO (español coloquial en América)*

14 de desembre

9.00 a 10.00	**Ponència. Martin Hilpert** (Université de Neuchâtel) *Construction Grammar and the analysis of spoken language*
10.00 a 10.30	Fien De Latte (Ghent University), Renata Enghels (Ghent University) & Linde Roels (Ghent University) *El lenguaje juvenil como catalizador del cambio lingüístico en el siglo 21: estudio contrastivo de los corpus COLAm y CORMA*
10.30 a 11.00	Ana Albalat Mascarell (Universitat Politècnica de València) & Maria Luisa Carrió (Universitat Politècnica de València) *La construcción de la intersubjetividad en el discurso político: Análisis de los marcadores de compromiso en la campaña electoral del 26-J*
11.00 a 12.30	Descans + pòsters
12.30 a 13.00	Marta Albelda Marco (Universitat de València) *Variación de la atenuación según el género discursivo: un estudio de corpus*
13:00 a 13.30	Carolina Figueras (Universitat de Barcelona) & M. Amparo Soler (Universitat de València) *Atenuación, ilocución e imagen*
13.30 a 15.00	Dinar
15.00 a 15.30	Romain Isely (ELCF, University of Geneva), Isabelle Racine (ELCF, University of Geneva), Sylvain Detey (Waseda University) & Julien Eychenne (Hankuk University of Foreign Studies) *Processing native and non-native speech corpus data: a phonological illustration with the French schwa*
15.30 a 16.00	Esteve Clua (Universitat Pompeu Fabra) & Miquel Salicrú (Universitat de Barcelona) *Aplicacions de la lògica difusa a l'anàlisi dels corpus dialectals: classificació de les varietats de frontera*
16.00 a 16.30	Descans
16.30 a 17.30	**Ponència**. **Xosé Luís Regueira** (Universidade de Santiago de Compostela) *Corilga: un corpus anotado multinivel para estudiar la variación y el cambio en la lengua hablada*
17.30 a 17.45	Clausura

Pòsters

14 de desembre

11.00 a 12.30	Noelia de La Torre Martínez (Universitat de València) *L'atenuació en el català i l'espanyol de València: anàlisi contrastiva*
	Irantzu Epelde (CNRS) *Dialect contact in a Basque valley*
	Hanna Jokela (University of Turku) & Milla Luodonpää-Manni (University of Turku) *Construing scientific knowledge and politeness in ten doctoral defenses held in Finnish and in French – an exploratory study of spoken academic discourse*
	Julia Kaiser (IDS Mannheim), Evi Schedl (IDS Mannheim) & Thomas Schmidt (IDS Mannheim) *Building up a multi-purpose reference corpus of spoken interactions*
	Juan Lorente Sánchez (Universidad of Málaga) *Give it me back on our wedding day: On the alternative double object construction on spoken Asian varieties of English*
	Nausica Marcos Miguel (Denison University) & Claudia Sánchez-Gutiérrez (University of California, Davis) *A Spanish Second Language Classroom Corpus: Discussing its Construction and Shareability*
	David Navarro Ciurana (Universitat de València) *El valor atenuante del discurso directo de pensamiento en el español de América*
	Anastasia Panova (National Research University Higher School of Economics, Moscou) & Ruprecht von Waldenfels (Friedrich Schiller University Jena) *A collection of non-standard spoken Russian corpora: approaches, tools and research*

Lloc / Venue

Sala de Graus Enric Valor (1a planta).
Facultat de Filologia, Traducció i Comunicació. Universitat de València.

Comitè científic

Marta Albelda (Universitat de València), Núria Alturo (Universitat de Barcelona), Dolores Azorín (Universitat d'Alacant), Miriam Bouzouita (Universiteit Gent), Josefina Carrera-Sabaté (Universitat de Barcelona), Ernestina Carrilho (Universidade de Lisboa), María-Luisa Carrió-Pastor (Universitat Politècnica de València), Emili Casanova (Universitat de València), Jaume Corbera (Universitat de les Illes Balears), Bert Cornillie (Katholieke Universiteit Leuven), Miquel Esplà (Universitat d'Alacant), Maria Estellés (Universitat de València), Inés Fernández-Ordóñez (Universidad Autónoma de Madrid), Elisa Fernández-Rei (Universidade de Santiago de Compostela), Mikel Forcada (Universitat d'Alacant), Mar Garachana (Universitat de Barcelona), Pedro Gras (Universiteit Antwerpen), Johannes Kabatek (Universität Zürich), Maria-Rosa Lloret (Universitat de Barcelona), Anke Lüdeling (Humboldt-Universität zu Berlin), Josep Martines (Universitat d'Alacant), Vicent Martines (Universitat d'Alacant), Brauli Montoya (Universitat d'Alacant), Sandra Montserrat (Universitat d'Alacant), Pere Navarro (Universitat Rovira i Virgili), Florent Perek (University of Birmingham), Manuel Pérez Saldanya (Universitat de València), Clàudia Pons-Moll (Universitat de Barcelona), Pilar Prieto (Universitat Pompeu Fabra), Claus Pusch (Universität Freiburg), Joan-Rafael Ramos (Universitat de València), Xosé-Luís Regueira-Fernández (Universidade de Santiago de Compostela), Josep Ribera (Universitat de València), Leonor Ruiz-Gurillo (Universitat d'Alacant), Pelegrí Sancho (Universitat de València), Andreu Sentí (Universitat de València), Juan-Andrés Villena-Ponsoda (Universidad de Málaga).

Comitè organitzador

Vicent Beltran (UA), Miquel Esplà (UA), Maribel Guardiola (UA), Jesús Jiménez (UV), Maria Josep Marin (UV), Sandra Montserrat (UA), Carles Segura (UA), Andreu Sentí (UV), Manuel Badal (UV), Paula Cruselles (UV), Pau Martín (UV), Caterina Martínez (UA), Anqi Tang (UV).

Resums de les conferències plenàries

Los corpus de conversaciones coloquiales. La elaboración del corpus AMERESCO (español coloquial en América)

Antonio Briz (Grupo Val.Es.Co., Universitat de València)

El panorama actual sobre corpus orales es, sin duda, alentador. Una revisión de los ya elaborados muestra una imagen rica y variada de estos, aunque todavía imperfecta.

Por un lado, siguen sin resolverse algunas cuestiones y problemas relacionados con la cantidad o calidad de los datos, la suficiencia de los corpus, los accesos a la información, la digitalización, los sistemas de marcación y de transcripción, la explotación, el trabajo de análisis y de abstracción… En esta conferencia intentamos plantear interrogantes sobre lo que se ha hecho y, especialmente, sobre lo que queda por hacer. Así, ¿son suficientes los corpus orales de que actualmente disponemos? ¿cuáles son las carencias más notables? ¿cómo deberían presentarse esos nuevos corpus? Y algo también muy importante, ¿ayudan nuestros análisis a partir de corpus a construir, confirmar, destruir o desconfirmar teorías?

Por otro lado, en mayoría de los grandes corpus, lo oral ocupa un espacio que no sobrepasa el 10% y dentro de estos hay una representación mínima de corpus de conversaciones (coloquiales). Ello muestra que, si bien se ha avanzado mucho, es preciso seguir elaborando corpus orales por el bien de los análisis. En 2012, lanzamos la propuesta de elaborar un macrocorpus representativo de la variedad situacional coloquial del español en España e Hispanoamérica y, en concreto, de conversaciones. Hoy ya es una realidad y muy especialmente el corpus de conversaciones coloquiales de América, AMERESCO. Los materiales, que pueden ya consultarse (http://esvaratenuacion.es/corpus-discursivo-propio/), están contribuyendo al estudio y comparación de la conversación coloquial en distintas ciudades de América. Ofrece información sobre este corpus es otro de los objetivos de esta exposición

Construction Grammar and the analysis of spoken language
Martin Hilpert (Université de Neuchâtel)

Up to now, spoken language remains a relatively blind spot of Construction Grammar. In my presentation, I will argue that constructional research on language has been affected by what has been called "the written language bias", and I will outline a number of ideas that can be pursued in order to reduce that bias. The main goal will be to develop a set of conceptual tools that allow us to conduct constructional analyses that take the inherent temporality of spoken language into account. I will specifically draw on Paul Hopper's theory of emergent grammar and on Peter Auer's framework of onlie-syntax, and I will discuss the notions of projection, expansion, and retraction. In order to show how these notions can be incorporated into constructional analyses, I will discuss two case studies, namely WH-clefts and a phenomenon that I call collaborative insubordination.

Anke Lüdeling (Humboldt-Universität zu Berlin)

Optionality and variation in spoken learner language

How do learners of a foreign language acquire variable and optional phenomena? How can a spoken corpus be used to study acquisition patterns? How can we analyze something that is not in the corpus?
I will exemplarily discuss these questions by analyzing disfluencies in spoken corpus of German as a Foreign Language. My focus is on methodological issues:

(a) multi-layer analysis and the need for different levels of normalization
(b) variationist models and the analysis of missing variants
(c) within-group variation.

Corilga: un corpus anotado multinivel para estudiar la variación y el cambio en la lengua hablada

CORILGA (Corpus Oral Informatizado da Lingua Galega) (http://ilg.usc.es/gl/proxectos/corpus-oral-informatizado-da-lingua-galega-corilga) (García-Mateo et al. 2014; Seara et al. 2016) es un corpus diseñado para posibilitar el estudio de la variación y el cambio lingüístico en el gallego oral. El corpus, todavía en una fase temprana de desarrollo, se compone de grabaciones de diferentes variedades de lengua, desde textos literarios (como recitados de poesía y representaciones teatrales) hasta conversaciones informales, pasando por textos de la actividad pública (conferencias, debates, intervenciones parlamentarias) y de los medios de comunicación (programas de televisión, informativos, doblaje de cine). Estas grabaciones abarcan desde mediados de la década de 1960 hasta la actualidad.

El buscador permite filtrar los resultados por el grado de formalidad y por el tipo de texto, de modo que se facilita el estudio de la variación. Además, también se pueden filtrar las búsquedas por año de grabación y por las características de los hablantes (edad y sexo, entre otras), por lo que resulta posible estudiar el cambio lingüístico tanto en tiempo aparente como en tiempo real.

El corpus contiene una transcripción ortográfica y una transcripción fonética realizadas en el programa Elan (Brugman & Russel 2004). Los audios están alineados a la anotación ortográfica, al nivel de la palabra y al segmento fonético, por medio de una herramienta de alineamiento automático. El programa dispone también de una herramienta de reconocimiento de habla y de transcripción automática, que actualmente está en fase de evaluación y de mejora. Se han incorporado igualmente herramientas de lematización y la adaptación al gallego del etiquetador morfológico Freeling (Guinovart & Solla 2017). De esta forma, se obtienen varias líneas (tiers), alineadas todas ellas al audio, que contienen la transcripción fonética, transcripción ortográfica, palabra, lema y etiquetas morfológicas.

En esta intervención, además de describir las características y el diseño de este corpus, se avanzarán ideas sobre sus posibilidades de aplicación y se comentarán algunos de los problemas y debilidades detectadas.

Referencias

Brugman, H & A. Russel (2004). Annotating Multimedia/ Multi-modal resources with ELAN. Proceedings. 4th International Language Resources and Evaluation Conference (LREC 2004). Lisboa, 2065-2068. http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf.

García-Mateo, C., A. Cardenal, X. L. Regueira, E. Fernández Rei, M. Martinez, R. Seara, R. Varela, N. Basanta (2014): CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis. Proceedings. 9th Language Resources and Evaluation Conference (LREC 2014). Reykjavik, 2653-2657. http://www.lrec-conf.org/proceedings/lrec2014/pdf/739_Paper.pdf

Guinovart, X. & M. Solla (2017): Freeling galego. Análise automática do galego. http://sli.uvigo.es/lingua/ (consulta 20.09.2018).

Seara, R., M. Martinez, R. Varela, C. García-Mateo, E. Fernández-Rei, X. L. Regueira (2016): Enhanced CORILGA: Introducing the Automatic Phonetic Alignment Tool for Continuous Speech. Proceedings. 10th Language Resources and Evaluation Conference (LREC 2016). Portorož, Slovenia, 2893-3898. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1074_Paper.pdf

Resums de les comunicacions

La construcción de la intersubjetividad en el discurso político: Análisis de los marcadores de compromiso en la campaña electoral del 26-J

Ana Albalat-Mascarell y María Luisa Carrió-Pastor

De acuerdo con el modelo de metadiscurso interpersonal elaborado por Hyland (2005), uno de los aspectos fundamentales para la construcción de la intersubjetividad es el compromiso, entendido como la dimensión comunicativa que abarca todos aquellos recursos lingüísticos y retóricos empleados por el hablante para involucrar a sus oyentes en el discurso. Se trata de mecanismos interaccionales tales como los pronombres referidos al oyente (p.ej., "tú", "ustedes", "nosotros" inclusivo), las expresiones directivas que comprometen a realizar acciones físicas o cognitivas, las expresiones interrogativas como estrategia dialógica, la apelación al conocimiento compartido indicado mediante fórmulas explícitas (p.ej., "como todos sabemos"), y las digresiones parentéticas que incluyen comentarios personales. Todas estas estrategias obedecen a un propósito de afiliación por el cual el hablante reconoce las expectativas de inclusión de su oyente al tiempo que guía retóricamente sus interpretaciones (Fu & Hyland, 2014; Hyland, 2005, 2010, 2017; Ilie, 2003; Jiang & Hyland, 2015, 2016; Mur-Dueñas, 2008, 2011; Suau-Jiménez, 2016). Este trabajo aborda el análisis de los citados marcadores de compromiso empleados con una finalidad retórica y persuasiva por los candidatos a la presidencia en los discursos de campaña para las elecciones generales de España de 2016. Los objetivos del estudio son los siguientes: en primer lugar, comparar el uso y distribución de los marcadores de compromiso en discursos pertenecientes a actos de campaña y en un debate electoral televisado; en segundo lugar, determinar si existe también cierta correlación entre el uso de dichos marcadores y la posición ideológica que ocupan sus hablantes dentro del espectro político. Nuestro estudio se basa en un corpus oral compuesto por las transcripciones de los discursos producidos por los candidatos de las cuatro principales fuerzas políticas españolas (esto es, PP, PSOE, Unidos Podemos y Cs), así como del único debate televisado en el que participaron todos los candidatos. En la metodología empleada para el análisis del corpus, se ha utilizado la herramienta "Metool" desarrollada específicamente para identificar y analizar categorías metadiscursivas. Los resultados del estudio indican que los políticos españoles se relacionan de manera distinta con la audiencia según se trate de un discurso electoral o un debate. Notamos asimismo diferencias en el modo de apelar al público y de incluirlo en el discurso entre candidatos de distinto signo ideológico.

Palabras clave:

Metadiscurso, Marcadores de compromiso, Pragmática, Corpus orales, Discurso político.

Referencias:

- Fu, X. & Hyland, K. (2014). "Interaction in two journalistic genres: a study of interactional metadiscourse". *English Text Construction*, 7(1): 122-144.
- Hyland, K. (2005). *Metadiscourse*. London, UK: Continuum.
- Hyland, K. (2010). Metadiscourse: mapping interactions in academic writing. *Nordic Journal of English Studies,* 9(2): 125-143.
- Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics,* 113: 16-29.
- Ilie, C. (2003). Discourse and metadiscourse in parliamentary debates. *Journal of Language and Politics,* 2(1): 71-92.
- Jiang, F., & Hyland, K. (2015). "The fact that': Stance nouns in disciplinary writing". *Discourse Studies*, 17(5), 529-550.
- Jiang, F., & Hyland, K. (2016). "Nouns and academic interactions: A neglected feature of metadiscourse". *Applied Linguistics,* 2016: 1-25.
- Mur-Dueñas, P. (2008). "Analysing engagement markers cross-culturally: The case of English and Spanish business management research articles". *English as an additional language in research publication and communication*: 197-214.
- Mur-Dueñas, P. (2011). "An intercultural analysis of metadiscourse features in research articles written in English and in Spanish". *Journal of Pragmatics*, 43: 3068-3079.
- Suau-Jiménez, F. (2016). "What can the discursive construction of stance and engagement voices in traveler forums and tourism promotional websites bring to a cultural, cross-generic and disciplinary view of interpersonality?" *Ibérica*, 31: 199-220.

Variación de la atenuación según el género discursivo: un estudio de corpus

Marta Albelda Marco
(Universitat de Valéncia)

Se presenta un estudio contrastivo de la atenuación pragmática en diversos géneros discursivos con el fin de mostrar la doble función de este fenómeno pragmático en la comunicación: su función retórica (al servicio de los fines retóricos de cada género discursivo) y su función social de gestión de la imagen (dirigida al cuidado de las imágenes de los hablantes). Partimos de la hipótesis de que de acuerdo con la finalidad más retórico-argumentativa o más socializadora de cada género discursivo, las categorías atenuantes empleadas variarán y se especializarán según el género (Markkanen y Schröder (eds.) 1997, Morales 2010, Contreras 2012, Villalba 2016, Cestero 2017).

Asimismo, se pretende demostrar que la atenuación se emplea por necesidades de protección o reparación de la imagen social, por lo que se considera que el cuidado de la imagen es una de los rasgos definidores de la atenuación. En consecuencia, la variación en el uso de la atenuación también puede indicar un distinto tratamiento de la imagen según el género del que se trate (Goffman 1965, Ho 1976, Schwartz y Bilsky 1990, Bravo 1999, Ting-Toomey y Oetzel 2003, Ho, Spencer-Oatey 2007, Hernández Flores 2013).

Entre las características que diferencian los géneros discursivos se encuentra un específico tratamiento de la imagen de los interlocutores, lo que viene determinado por los propósitos comunicativos del género y por la naturaleza de participantes que intervienen. Atender y ajustarse a las expectativas de imagen de cada género es un requisito más del éxito comunicativo. Así pues, para llevar a cabo este estudio se ha realizado un análisis exploratorio de un corpus de cinco géneros recopilado *ad hoc*: artículos de investigación, foros de comunicación en línea, conversaciones coloquiales, mesas redondas de discusión de expertos y debates políticos. Estos géneros presentan rasgos en contraste en el canal de comunicación, en el carácter dialogal/monologal, en la repercusión social que presentan, en los rasgos situacionales y en la función comunicativa primordial del género.

Los resultados del estudio confirman la variación formal y funcional de la atenuación y muestran cómo los mecanismos atenuantes tienden a especializarse en un tipo de tratamiento de la imagen social, y en consecuencia en función del género discursivo.

Bibliografía

Bravo, Diana. 1999. "¿Imagen positiva vs. imagen negativa? Pragmática social y componentes del face". *Oralia* 2: 155-184.

Cestero, Ana María. 2017. "La atenuación en el habla de Madrid: patrones sociopragmáticos". *Rilce* 33.1: 57-86.

Contreras, Josefa. 2012. "¿Hay diferencia en las estrategias de atenuación en los correos-electrónicos españoles y alemanes?". *Oralia* 15: 325-242.

Goffman, Erving. 1967. *Interaction Ritual: Essays in Face-to-Face Behavior*. Chicago: Aldine.

Hernández-Flores, Nieves. 2013. "Actividad de imagen. Caracterización y tipología en la interacción comunicativa". *Pragmática Sociocultural* 1(2): 1–24.

Ho, David Yau-Fai. 1976. "On the Concept of Face". *American Journal of Sociology*, 81(4): 867–84.

Markkanen, Raija y Hartmut Schröder (eds.). 1997: *Hedging and Discourse. Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Berlín: Walter de Gruyter.

Morales, Óscar. A. 2010: *Los géneros escritos de la Odontología hispanoamericana. Estructura retórica y estrategias de atenuación en artículos de investigación, casos clínicos y artículos de investigación*. Barcelona: UPF.

Schwartz, Shalom H., and Wolfang Bilsky. 1990. "Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications". *Journal of Personality and Social Psychology* 58(5): 878-891.

Spencer-Oatey, Helen, 2007. "Theories of identity and the analysis of face". *Journal of Pragmatics* 39: 639–656.

Ting-Toomey, Stella, and John G. Oetzel. 2003. "Face concerns in interpersonal conflict". *Communication Research* 30(6): 599-624.

Villalba, Cristina (2016): *Actividades de imagen, atenuación e impersonalidad en los juicios orales*. Tesis Doctoral. Universitat de València.

A corpus of informal speech for Valencian Catalan: *Parlars*

Vicent Beltran, Miquel Esplà, Maribel Guardiola, Sandra Montserrat, Carles Segura
(Universitat d'Alacant) & Andreu Sentí (Universitat de València)

Catalan has various textual corpora such as the CTILC2 and the CIVAL which contain texts of the contemporary written language. Although we are fortunate to have spoken corpora and speech data, such as the CCCUB corpus (Alturo et al. 2004; Boix-Fuster et al. 2007; Carrera-Sabaté & Vilaplana; Payrató & Alturo 2002; Pons & Vilaplana 2009; Vilaplana & Perea 2003; Vilaplana et al. 2007) or the *Atles entonatiu* (Prieto & Cabré 2007-2012), there is still a gap in the materials available. We need a large textual corpus in order to study variation in the colloquial oral language (phonetics, morphosyntax, semantics, lexicon and pragmatics) (Fernández Ordóñez 2011; cf. Llop & Pineda 2017). Furthermore, there is also the fact that there is no corpus dedicated to informal conversational Valencian Catalan. For this reason, we have begun to construct the *Corpus Parlars. Corpus oral del valencià col·loquial*, which has the following objectives:

1. To document the less corrupted informal Catalan possible (Beltran & Segura-Llopes 2017), before it disappears in the face of the pressure exerted by the standard form of the language and, above all, Spanish, which is diluting the language at a great rate throughout the districts where it is spoken (Segura 2003).
2. To provide researchers with suitable materials to carry out descriptive and analytical studies of the linguistic variation in Catalan, especially functional (colloquial) and dialectal.
3. To test the hypothesis that languages function as conventionalized structures, as described by Cognitive Constructions Grammar (Goldberg 2003; Taylor 2012; Hilpert 2014).

The focus on variation in the current language conditions the type of corpus that we are creating. In the present article we present the initial phase in the creation of the *Corpus Parlars*, which consists of precisely defining its principal characteristics and working methodology. First we will develop the objectives, which we have already described, and the design of the corpus, that is, which types of text represent the colloquial variety and which types of speakers (informants) have the purest form of Valencian. We will then focus on the working methodology; that is, data collection, data handling, query interface.

Regarding data collection, we will establish which localities will be chosen for the sample, the profile of the informants (age, socioeconomic status, education), the type of interaction (spontaneous conversation, semi-directed conversation, monologue-narrative), duration of the interviews, etc. In terms of data handling, we will present the criteria that we are following for the transcription and compare them with other current corpora (Payrató & Alturo 2002; Briz 2002; Hidalgo & Sanmartín 2005; Bladas 2009), the type of tagging and the standards for the morphosyntactic annotation of the corpus (Ide & Pustejovsky 2017) and the natural language processing technologies used to support the transcription and annotation process. Finally, we will explain the options for asking the corpus questions via the query interface.

Referencias bibliográficas

Alturo, Núria; Òscar Bladas, Marta Payà y Lluís Payrató (ed.) (2004): *Corpus oral de registres. Materials de treball.* Barcelona: Publicacions y Edicions de la Universitat de Barcelona.

Beltran, Vicent; Segura-Llopes, Carles (2017): Els *parlars valencians*. València: PUV.

Bladas, Òscar (2009): *Manual de transcripció del discurs oral. Materials de treball*. Universitat de Barcelona.

Boix-Fuster, Emili; Marina Àlamo Sala, Mireia Galindo Solé, Francesc Xavier Vila y Moreno (ed.) (2007): *Corpus de Varietats Socials. Materials de treball*. Barcelona: Publicacions i Edicions de la Universitat de Barcelona.

Briz Gómez, Antonio y Grupo Val.Es.Co. (2002): "Corpus de conversaciones coloquiales", Anejo de la revista Oralia, Madrid, Arco-Libros.

Carrera-Sabaté, Josefina y Joaquim Viaplana (ed.): *Corpus Oral Dialectal (COD). Textos orals del nord-occidental*. Dipòsit Digital de la UB.

CCCUB = Corpus del català contemporani de la Universitat de Barcelona. [http://www.ub.edu/cccub/]

CIVAL = Acadèmia Valenciana de la Llengua: *Corpus Informatitzat del Valencià* [http://cival.avl.gva.es/]

CTILC2 = Institut d'Estudis Catalans: *Corpus textual informatitzat de la llengua catalana* [https://ctilc2.iec.cat/]

Fernández Ordóñez, Inés (2011): "Nuevos horizontes en el estudio de la variación gramatical del español: el Corpus Oral y Sonoro del Español Rural", en G. Colón & Ll. Gimeno (eds.), *Noves tendències en la dialectologia contemporània*, Castelló de la Plana, Universitat Jaume I, págs. 173-203.

Goldberg, A. E. (2003): "Constructions: A new theoretical approach to language", Trends in Cognitive Sciences 2 (5), 219-224.

Hidalgo, A. & J. Sanmartín, (2005): «Los sistemas de trascripción de la lengua hablada», *Oralia*, 8, pp. 13-36.

Hilpert, M. (2014): *Construction Grammar and its Application to English*. Edimburgh, Edimburgh University Press.

Ide, Nancy & James Pustejovsky (eds.) (2017): *Handbook of Linguistic Annotation*. Dordrecht: Springer.

Llop, Ares & Anna Pineda (2017) "L'estudi de la variació sintàctica en català: On som i cap on anem?". En Manuel Pérez Saldanya & Rafael Roca i Ricart (ed.): *Actes del XVIIè Col·loqui Internacional de Llengua i Literatura Catalanes. Universitat de València, 7-10 de juliol de 2015*. Barcelona: IEC, p. 527-542.

Payrató, Lluís y Núria Alturo (ed.) (2002): *Corpus oral de conversa col·loquial. Materials de treball.* Barcelona: Publicacions de la Universitat de Barcelona.

Perea, Maria-Pilar y Joaquim Viaplana: *Corpus Oral Dialectal (COD)*. *Selecció de textos*. Dipòsit Digital de la UB.

Pons, Clàudia y Joaquim Viaplana (ed.) (2009): *Corpus oral dialectal (COD)*. *Textos orals del balear*. Dipòsit Digital de la UB.

Prieto, Pilar & Cabré, Teresa (coords.) (2007-2012): *Atles interactiu de l'entonació del català*. Pàgina web: <http://prosodia.upf.edu/atlesentonacio/>.

Segura, Carles (2003): *Variació dialectal i estandardització al Baix Vinalopó.* Alacant / Barcelona: Institut Interuniversitari de Filologia Valenciana / Publicacions de l'Abadia de Montserrat.

Taylor, John R. (2012): *The mental corpus*. Oxford: Oxford University Press.

Viaplana, Joaquim y Maria Pilar Perea (ed.) (2003): *Textos orals dialectals del català sincronitzats. Una selecció*. Barcelona: Promociones y Publicaciones Universitarias (PPU).

Viaplana, Joaquim; Maria-Rosa Lloret, Maria-Pilar Perea y Esteve Clua (2007): *COD*. *Corpus Oral Dialectal*. Barcelona: Promociones y Publicaciones Universitarias (PPU).

Methods, resources and some results of the C-ORAL-BRASIL project for Brazilian Portuguese spoken corpora

Giulia Bossaglia[1]
Lúcia de Almeida Ferrari[2]

In this talk we will present in detail a few aspects of the C-ORAL-BRASIL project (Raso & Mello 2012), aimed at the compilation of Brazilian Portuguese (BP) spoken corpora. The project was born as an out-of-Europe branch of the C-ORAL-ROM (Cresti & Moneglia 2005), which has compiled spoken corpora for Italian, French, Spanish and European Portuguese. So far, the C-ORAL-BRASIL has compiled four corpora: Informal (published in 2012), Formal, Media and Telephone (to be published soon).

Like the other C-ORAL corpora, the C-ORAL-BRASIL can be considered an appropriate tool for the analysis of spoken language, since it permits access not only to the transcripts of the recording sessions (with prosodic annotation of terminal and non-terminal breaks), but also to the audio files and the text-to-speech alignment (through the WinPitch software: Martin 2015). Internal concordance between transcribers and prosodic boundaries' annotators was statistically validated through the Kappa test (Fleiss 1971; Raso & Mittmann 2009).

The C-ORAL-BRASIL corpora are representative of the Minas Gerais State diatopy of spoken BP, mostly from the capital city Belo Horizonte – other diatopic varieties appear at times as well. The aim of the corpora is to document *spontaneous*, non-elicited speech, and non-invasive recording equipment (lapel mics) was used to minimize interferences over the naturalness of the recordings. The metadata (sociolinguistic information about the speakers, context of the recording, and so on) of each recording session are made available as well.

All corpora were compiled to achieve the greatest diaphasic variation of spontaneous interactions: a wide range of contexts for recording sessions is attested in the Informal and Formal ones, a variety of TV and radio formats makes up the Media one, and informal conversations on the most diverse topics are found in the Telephone corpus.

Besides the four corpora, two informationally tagged minicorpora were compiled within the C-ORAL-BRASIL project: the BP one (Panunzi & Mittmann 2014), a representative sample of the C-ORAL-BRASIL Informal corpus, and the American English one, sampled and adapted to the C-ORAL transcription criteria (Cavalcante & Ramos, 2016) from the *Santa Barbara Corpus of Spoken American English* (Du Bois et al. 2000-2005). The BP minicorpus was compiled to boost cross-linguistic comparison with an already existing Italian minicorpus, sampled from the Italian C-ORAL-ROM Informal corpus (the DB-IPIC minicorpus: Panunzi & Mittmann 2014); the American English minicorpus allows a valuable cross-linguistic comparison outside the Romance languages' domain. All three minicorpora are comparable in size and architecture: as the

[1] Universidade Federal de Minas Gerais, Belo Horizonte (Brazil).
[2] Universidade Federal de Minas Gerais, Belo Horizonte (Brazil).

reference corpora from which they were extracted, they are provided with transcripts, metadata, audio files and text-to-speech alignment, PoS tagging (at the moment only for Italian and BP), and information tagging.

The theoretical basis for the information tagging stems from the *Language into Act Theory* (Cresti 2000; Moneglia & Raso 2014), an extension of Austin's (1962) Speech Act Theory that emphasizes the importance of prosody for the structuring of spoken language. Information units' tagging was made manually by a team of trained annotators – a thorough revision of the tagging has been completed lately by a new team.

Due to the information tagging, the minicorpora are relevant resources for the cross-linguistic study of several aspects of spoken language, especially at the interface with information structure. Therefore, a brief overview of published studies and ongoing researches based on these resources will be provided.

References

Cavalcante, F. A., & Ramos, A. C. (2016). The American English spontaneous speech minicorpus. Architecture and comparability. *CHIMERA: Romance Corpora and Linguistic Studies*, *3*(2), 99-124.

Cresti, E. (2000). *Corpus di italiano parlato: Introduzione* (Vol. 1). Accademia della Crusca.

Cresti, E. & Moneglia, M. (Eds.). (2005). *C-ORAL-ROM: integrated reference corpora for spoken romance languages* (Vol. 15). John Benjamins Publishing.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000). Santa Barbara Corpus of Spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Martin, P. (2015). WinPitch. 2011.

Moneglia, M., & Raso, T. (2014). Notes on Language into Act Theory. *Raso, Tommaso/Mello, Heliana (Eds.), Spoken Corpora and Linguistic Studies. Amsterdam-Philadelphia, Benjamins*, 468-495.

Raso, T. & Mello, H. (Eds.). (2012) *C-oral-Brasil: corpus de referência do português brasileiro falado informal. I.* Editora UFMG.

Raso, T., & Mittmann, M. M. (2009). Validação estatística dos critérios de segmentação da fala espontânea no corpus C-ORAL-BRASIL. *Revista de Estudos da Linguagem*, *17*(2), 73-91.

In Search of New Transcription and Annotation Methods:
The Collaborative Game-Based Approach

Miriam Bouzouita
Ghent University

In comparison to other European languages, the morphosyntax of the Spanish dialects remains little explored, despite the recent surge in interest in dialect grammars, not only from dialectologists but also formal linguists and typologists (e.g. Fernández-Ordóñez 2012, Gallego in press [2018], among others). In our view, one of the reasons for this situation is the lack of large annotated dialectal corpora. In this talk, we will present a newly funded project which aims to fill this lacuna by creating a morphosyntactically annotated and parsed corpus of the European Spanish dialects. This dialect corpus is being designed in a geographically balanced way and its material proceed from *the Corpus Oral y Sonoro del Español Rural* corpus (COSER corpus; the 'Audible Corpus of Spoken Rural Spanish' corpus). The COSER corpus is the largest collection of oral data in the Spanish-speaking world: at the moment it contains more than 1700 hours of recorded material from semi-directed interviews with a total of 2476 informants, of which 47.6% (1180/2476) are men and 52.3% (1296/2476) women. These informants are all elderly speakers (average age: 72.9 years), who enjoyed (very) limited education opportunities. As concerns the geographic coverage of the COSER corpus, since March 2018, dialectal material from 50 provinces and islands has been gathered.

One mayor problem of the COSER corpus, however, is that most of the dialect recordings remain untranscribed: recent calculations indicate that around 40% of the gathered dialect data is transcribed, of which only 207 hours have been revised and are publicly available through the corpus website (www.corpusrural.es). As transcribing and annotating are expensive and labour-intensive, this new project is exploring new methods for obtaining transcriptions and (morphosyntactic) annotations. More specifically, the collaborative gamification approach is being tested in order to construct the parsed corpus of European Spanish dialects. In other words, a crowdsourced game is being created through which members of the general public contribute to the co-creation of the parsed corpus by providing transcriptions and annotations in the context of a so-called serious game or a Game With A Purpose (e.g. von Ahn 2006).

Fernández-Ordóñez, Inés (2012) 'Dialect Areas & Linguistic Change: Pronominal Paradigms in Ibero-Romance'. In Vogelaer, G. & G. Seiler (eds), *The Dialect Laboratory.* Amsterdam: Benjamins, 73-106.

Gallego, Ángel (ed., in press [2018]) *Syntactic Variation in Spanish Dialects.* Oxford: Oxford University Press.

von Ahn, Luis (2006) Games with a purpose. *Computer* 39.6, 92-94.

A parsed corpus of Southern Dutch dialects

Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen & Jacques Van Keumeulen
Ghent University

This paper reports on the creation of an annotated corpus of spontaneous Southern Dutch dialect speech. It is based on a collection of 783 recordings (about 700 hours) from 617 locations in the Dutch-speaking provinces in Belgium, Zealand Flanders (Netherlands) and French Flanders (France). The tapes were recorded in the 1960s and 1970s, and contain the speech of dialect speakers born around the turn of the 20th century, the oldest informant being born in 1871. The dialect is hardly influenced by the standard language, as the speakers are almost exclusively monolingual and received only minimal formal education. The collection is therefore of immense value for linguistic research. Furthermore, it is a unique historical and cultural-historical resource, as it reports on topics such as both World Wars, the rise of electricity, bikes and cars and offers a unique perspective on lost professions and leisure activities. Although the recordings are already digitized and available online, they have not been transcribed or annotated systematically, which makes them hardly searchable for word forms or content, let alone for structures. The digital markup and exploration of this valuable treasure is an urgent desideratum considering the rapid dialect loss in Flanders, which means that soon there will be no one able any longer to transcribe the recordings.

The paper reports more specifically on an ongoing pilot project to transcribe and linguistically annotate about 40 recordings in the preparation of a larger infrastructure project to eventually make all the recordings accessible for fundamental research. We will first outline the transcription protocol developed specifically for this project, and then focus on the annotation pipeline, which starts with time-aligned transcriptions in ELAN (https://tla.mpi.nl/tools/tla-tools/elan/). The transcriptions consist of two layers, one closer to the dialect (cf. 1a, for instance capturing clitic constructions – marked with # – and ablaut phenomena), and one closer to Standard Dutch (2a), in order to make the data more searchable.

(1)	a	neen#t…	k#en	ik	ewrocht	met	een		ploef
	b	neen het...	ik heb	ik	gewrocht	met	een		ploeg
		no it…	I have	I	workedwith	a		plough	
		‘No (that is not the case), I have worked with a plough.’							

The time-alignment between audio and transcription facilitates (among others) phonetic research (as the transcription itself is not phonetic). In order to improve the quality of the transcriptions, we

make use of crowdsourcing by setting up a network of dialect-speaking volunteers to check the transcriptions and to resolve ambiguities or doubts. After the transcription phase, the data is tokenized, lemmatized, PoS-tagged and parsed. We opt for an enrichment of ELAN-xml, as this allows maintaining the association with the time codes/the audio. The PoS-tags are awarded automatically and corrected manually. The tagset is still under development, but will connect as much as possible with existing tagsets for Dutch such as the CGN tagset in order to facilitate large-scale comparative research. The same considerations of interoperability guide the syntactic annotation, which follows the format of the Penn parsed corpora of historical English. The parsing is partly automated as well, using a pipeline of scripts for revision and shallow parsing, amongst others with CorpusSearch revision queries and the graphical user interface Annotald (https://annotald.github.io). Finally, it is the intention to combine audio, aligned transcriptions and annotations in a sustainable and searchable online corpus, made available via CLARIN in collaboration with the *Instituut voor Nederlandse Taal* (INT). At a later stage, the digital transcriptions can be subject to topic modeling, as such creating infrastructure for historical research.

Dificultades y complejidades en el diseño de un entorno web de consulta: el caso del corpus oral Ameresco

Adrián Cabedo Nebot
Adrian.cabedo@uv.es
Universitat de València – Grupo Val.Es.Co

En esta comunicación pretendemos analizar las dificultades experimentadas en la configuración de un corpus textual (Rojo Sánchez, 2010, 2015, 2017) y, más concretamente, de un corpus conversacional de español hablado (Love, Dembry, Hardie, Brezina, & McEnery, 2017). En tal sentido, se observan ya de entrada problemas evidentes, puesto que los corpus orales existentes de español suelen centrarse en la variante peninsular (Cabedo & Pons, 2013), o bien recogen formatos interactivos no conversacionales, como entrevistas (Cestero Mancera, 2012; Fernández-Ordóñez, 2005) o programas radiofónicos y televisivos. No obstante, sí existen algunos modelos de corpus conversacionales, aunque para grupos etarios concretos (Jorgensen, 2007). Los principales problemas, así mismo, se relacionan con el sistema de transcripción y de etiquetas utilizado, por un lado, y, también, con el formato de los datos (XML, JSON, texto plano, documentos tabulares como Excel, etc.) o con los programas informáticos utilizados (ELAN [Max Planck Institute, 2017], PRAAT [Boersma & Weenink, 2017], etc.).

Como novedad en ese marco de corpus conversacionales de español, presentaremos la plataforma en construcción del corpus oral Ameresco (Albelda Marco & Estellés Arguedas, 2017), que recoge conversaciones de diferentes variantes geográficas de español hablado: Iquique (Chile), Habana (Cuba), Tucumán (Argentina), Valencia (España), México DF (México), Barranquilla (Colombia)... Este corpus, todavía en proceso de recolección y ampliación, contiene en la actualidad 43 conversaciones espontáneas.

Así pues, el interés principal es mostrar su proceso constitutivo desde la transcripción mediante un sistema de etiquetado y con el uso del programa ELAN (Max Planck Institute, 2017) hasta la implementación en un entorno web de consulta mediante una base de datos PostgreSql (www.esvaratenuacion.es/corpus).

Finalmente, más allá de los inconvenientes observados, esta comunicación también tiene como objetivo realizar una reflexión crítica sobre la viabilidad de un entorno de consulta que sea adaptable al mayor número de intereses científicos. Sobre todo, se problematiza sobre el formato en que deben presentarse los datos y sobre la agilidad y flexibilidad de labores como la transcripción, el etiquetado morfosintáctico o el análisis melódico de los archivos de audio incorporados al corpus.

Bibliografía

Albelda Marco, M., & Estellés Arguedas, M. (2017). *Corpus Ameresco*. Retrieved from www.esvaratenuacion.es

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/

Cabedo, A., & Pons, S. (2013). Corpus Val.Es.Co 2.0. Retrieved from www.valesco.es

Cestero Mancera, A. M. (2012). El Proyecto para el estudio sociolingüístico del español

de España y América (PRESEEA). *Español Actual: Revista de Español Vivo*, (98), 227–236.

Fernández-Ordóñez, I. (2005). *El Corpus Oral y Sonoro del Español Rural*. Cantoblanco: Universidad Autónoma de Madrid, 2005.

Jorgensen, A. (2007). COLA. Un corpus oral de lenguaje adolescente. In *Discurso y oralidad: homenaje al profesor José Jesús de Bustos Tovar* (Vol. 1, pp. 225–234). Madrid : Arco Libros, [2007]. Retrieved from https://dialnet.unirioja.es/servlet/extart?codigo=2549155

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.22.3.02lov

Max Planck Institute. (2017). ELAN (Version 5.2). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Rojo Sánchez, G. (2010). Sobre codificación y explotación de corpus textuales: Otra comparación del Corpus del español con el CORDE y el CREA. *Lingüística*, (24), 11–50.

Rojo Sánchez, G. (2015). Corpus textuales del español. In *Enciclopedia de Lingüística Hispánica* (Vol. 2, pp. 285–296).

Rojo Sánchez, G. (2017). Sobre la configuración estadística de los corpus textuales. *Lingüística*, *33*(1), 121–134.

Aplicacions de la lògica difusa a l'anàlisi dels corpus dialectals: classificació de les varietats de frontera.

Esteve Clua
Universitat Pompeu Fabra
esteve.clua@upf.edu

Miquel Salicrú
Universitat de Barcelona
msalicru@ub.edu

En dialectometria l'interès se centra a identificar i caracteritzar les varietats dialectals, interpretar les diferències espacials, i si cap, estudiar-ne l'evolució en el temps. En aquest context, l'abast (el corpus i l'extensió geogràfica) l'estableix l'interès lingüístic i les possibilitats del treball; la tècnica estadística i les eines computacionals permeten el tractament de la informació: classificació en grups i representació cartogràfica fonamentalment.

Amb independència de la mesura de distància utilitzada, la classificació de poblacions per mètodes deterministes (UPGMA, Ward i K-means, entre uns altres) proporciona un resultat constituït per grups disjunts, estancs entre si. La simplicitat en l'agrupació de poblacions que s'obté amb aquests mètodes té com a contrapartida algunes limitacions. Per superar aquestes limitacions, la lògica difusa (Fuzzy Logic), com per exemple, la classificació amb l'algorisme Fuzzy C-means, ha proporcionat resultats molt interessants en diferents àmbits de coneixement, ja que permet obtenir diferents graus de pertinença de les entitats a cadascun dels grups.

En aquesta comunicació presentem una anàlisi millorada de la distància lingüística de les varietats del català a partir de l'aplicació de mètodes Fuzzy a les dades del COD (Corpus oral dialectal del Català contemporani). El nostre objectiu és determinar de manera més adequada la classificació de les poblacions de frontera que expliquen el contínuum dialectal català i que se situen en les interfícies dels diferents grups dialectals d'aquesta llengua.

La proposta metodològica que defensem permet analitzar amb detall l'adscripció dialectal de determinades varietats del català que, per les seves característiques lingüístiques, han presentat un cert grau de complexitat a l'hora de ser classificades. Ens referim concretament a les varietats de la zona nord de Castelló, a les varietats tortosines i a les de l'anomenada Franja de Ponent.

La utilització de l'algorisme Fuzzy C-means permet, a més a més, analitzar amb facilitat la distància lingüística entre grups dialectals. D'altra banda, aquest algorisme, en definir graus

de pertinença de les varietats als diferents grups, facilita la percepció del canvi lingüístic i de fenòmens relacionats amb l'advergència o la divergència lingüística.

Referències bibliogràfiques

Clua, Esteve; Goebl, Hans; Casassas, Xavier; Civit, Sergi & Miquel Salicru (2013): "Anàlisi dialectomètrica del COD amb el suport del VDM", a Clua, Esteve & Maria Rosa Lloret (2013): Qüestions de morfologia flexiva i lèxica del català. Volum d'homenatge a Joaquim Viaplana. Colecció Symposia Philologica, 24. Alacant: Institut Interuniversitari de Filologia Valenciana, pàg. 133-168.

Valls, Esteve; Wieling, Martijn; Nerbonne, John (2013): «Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects». Literary and Linguistic Computing, 28(1), pàg. 119-146.

Viaplana, Joaquim; Lloret, Maria-Rosa, Perea, Maria-Pilar & Clua, Esteve (2007): COD. Corpus Oral Dialectal. Barcelona: PPU. (Publicació en CD- ROM).

El lenguaje juvenil como catalizador del cambio lingüístico en el siglo 21: estudio contrastivo de los corpus COLAm y CORMA.

De Latte, Fien; Enghels, Renata; Roels, Linde (Universidad de Gante)

En las últimas décadas, el lenguaje juvenil se ha convertido en uno de los campos de investigación preferidos en la sociolingüística, no solo porque es una variante en la que las intervenciones normativas influyen mucho menos, sino ante todo porque actúa como catalizador del cambio lingüístico, que prepara y señala nuevas tendencias que luego se difunden en la lengua corriente (Jørgenson 2008; Palacios Martínez y Núñez Pertejo 2014; Stenström 2014, Zimmerman 2002). Efectivamente, gracias a su rol clave en la cultura general (cf. su influencia en los medios de comunicación), los jóvenes han adquirido más prestigio como grupo social y se han convertido en modelos de conducta para otros grupos (Briz Gómez 2003). Este prestigio explica por qué, de su parte, muchos adultos buscan 'rejuvenecerse' mediante la imitación de expresiones lingüísticas tomadas del lenguaje adolescente. Esta actitud refuerza la alta velocidad con la que se introducen los cambios en el lenguaje juvenil. Es decir, para salvaguardar el símbolo de su identidad de grupo, los jóvenes se sienten estimulados a reaccionar contra la imitación y recurren a nuevas formas lingüísticas, generando así un movimiento cíclico de acción-reacción-acción (Briz 2003). Sin embargo, pocos estudios han 'monitorizado' sistemáticamente la velocidad con la cual innovaciones lingüísticas se introducen en el lenguaje juvenil.

Nuestra ponencia tiene un doble objetivo. Por un lado, aspira a dar a conocer el corpus CORMA (*Corpus Oral de Madrid*), que aporta nuevas dimensiones al estudio del español coloquial cotidiano. Por otro lado, quiere ilustrar su aptitud para el estudio de cambios lingüísticos recientes.

CORMA, grabado en 2016, recopila 59 conversaciones de tono informal entre hablantes madrileños, realizadas en diferentes ámbitos comunicativos. Incluye 43 horas grabadas, más de 300.000 palabras transcritas, y representa el habla de 391 hablantes, masculinos y femeninos, de todas las generaciones y de diferentes clases sociales. Las conversaciones se caracterizan por la igualdad entre los participantes, el ámbito discursivo familiar, la temática no especializada, la ausencia de planificación, y la finalidad interpersonal del acto comunicativo, de manera que se definen como conversaciones coloquiales prototípicas (Briz Gómez 2001: 51).

El estudio de caso analiza dos fenómenos típicos del lenguaje juvenil, a saber el uso excesivo de voces intensificadas (cf. *mazo, mogollón, que lo flipas* etc.) (Briz Gómez 1997; Martínez López 2009) y la alta productividad de expresiones vocativas (cf. *tío/tía*, *tronco/tronca*, o *chaval/chavala*) (Jørgenson 2008, 2013; Stenström 2008). Más en concreto quiere comparar el ritmo con el cual estos fenómenos se renuevan en el lenguaje juvenil. Contrastamos los datos del corpus COLAm (*Corpus Oral de Lenguaje Adolescente de Madrid*), recopilado a principios del siglo 21, con los datos de CORMA, grabado una década más tarde. Los primeros resultados muestran que las formas intensificadas parecen renovarse más rápidamente que las formas de tratamiento. Para dar unos ejemplos: *mazo*, el intensificador más productivo en COLAm, parece haber caído en desuso en el lenguaje juvenil actual tal y como se documenta en CORMA. En cambio, contrariamente a las expectativas, el uso del vocativo *tío*, ya omnipresente en COLAm, se ha extendido al español de otras generaciones en CORMA, sin que desaparezca en la variante juvenil. Estos datos sugieren que la velocidad con la cual cambios lingüísticos se introducen en el lenguaje juvenil depende del fenómeno lingüístico estudiado.

Referencias

Briz Gómez, A. (1997), "Los intensificadores en la conversación coloquial", en Briz Gómez, Antonio, José Ramón Gómez Molina, María José Martínez Alcalde & Grupo VAL.ES.CO (eds.), *Pragmática y gramática del español hablado: Actas del II Simposio sobre Análisis del Discurso Oral*. Valencia: Pórtico Libros, 13-36.

Briz Gómez, A. (2001), *El español coloquial en la conversación. Esbozo de pragmagramática*. Barcelona: Ariel.

Briz Gómez, A. (2003), "La interacción entre jóvenes. Español coloquial, argot y lenguaje juvenil", *Lexicografía y Lexicología en Europa y América*. Madrid: Gredos, 141-154.

Jørgenson, A. M. (2008), "*Tío* y *tía* como marcadores en el lenguaje juvenil de Madrid". en: Olza Moreno, I., Casado Valverde, M. & González Ruiz, R. (eds.): *Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL)*. Pamplona: Universidad de Navarra, 387-396.

Jørgenson, A. M. (2013), "Spanish teenage language and the COLAm corpus", *Bergen Language and Linguistics Studies* 3/1, 151-166.

Martínez López, Juan Antonio (2009), "Lexical innovations in Madrid's teenage talk: Some intensifiers", en Stenström, Anna-Brita & Annette Myre Jørgensen (eds.), *Youngspeak in a Multilingual Perspective*. Amsterdam: Benjamins, 81-93.

Palacios Martínez, Ignacio M. & Paloma Núñez Pertejo (2014), "Strategies used by English and Spanish teenagers to intensify language. A contrastive corpus-based study", *Spanish in Context*, 11(2):175-201.

Stenström, Anna-Brita (2008), "Algunos rasgos característicos del habla de contacto en el lenguaje de adolescentes en Madrid", *Oralia*, 11:207-226.

Stenström, Anna-Brita (2014), *Teenage Talk: From General Characteristics to the Use of Pragmatic Markers in a Contrastive Perspective*. Palgrave: Macmillan.

Zimmerman, Klaus (2002), "La variedad juvenil y la interacción verbal entre jóvenes", en Rodríguez González, Félix (ed.), *El lenguaje de los jóvenes.* Barcelona: Ariel, 137-164.

Problemas y soluciones en el diseño y construcción del corpus Ameresco

María Estellés Arguedas
Marta Albelda Marco
(Universitat de Valéncia)

El diseño, recolección y construcción del corpus Ameresco (América Español Coloquial) es uno de los objetivos principales del proyecto Es.VaG.Atenuación. Desde el año 2014 alrededor de quince equipos de diferentes universidades hispanoamericanas y españolas estamos trabajando de forma coordinada y bajo una metodología común para la creación de un corpus natural de conversaciones coloquiales. Entre otros, están en marcha y parcialmente recogidos, los corpus de Monterrey, Ciudad de México, La Habana, Panamá, Tucumán, Medellín, Barranquilla, Iquique y Valencia.

Los presupuestos metodológicos y de diseño del corpus Ameresco siguen en sus bases y fundamentos a los del corpus Val.Es.Co., recogido por primera vez en 1995 (Briz et alii 1995), con sucesivas ampliaciones (en 2002, Briz y Grupo Valesco; en línea, Pons y Cabedo, www.valesco.es). Se trata de un corpus de lengua hablada con muestras, por un lado, obtenidas en los espacios naturales donde tienen lugar, de forma secreta, con observador no participante. Por otro lado, están recogidas siguiendo un criterio de representatividad sociolingüística y transcritas de acuerdo con un sistema de símbolos diseñado para representar de la manera más fiel posible las particularidades de la oralidad conversacional.

Partiendo de esta base común, el corpus Ameresco, se propone construir un corpus más amplio en el que se representen las diversas variedades dialectales del español. La construcción de un corpus de variación y el desafío de convertir en accesible para la comunidad científica el uso y manejo de este corpus tanto en su versión oral como transcrita, ha mostrado una serie de problemas en distintos niveles del proceso de creación. La presente comunicación se propone, por tanto, señalar los principales problemas surgidos y las propuestas de solución que se han estudiado, indicando sus ventajas.

Así, los objetivos concretos de esta comunicación son: (i) caracterizar y presentar las bases teóricas y metodológicas del corpus Ameresco; (ii) discutir los principales problemas en la fase de recogida de datos (legalidad, calidad de los materiales, etc.) y en lo relativo a la representatividad sociolingüística; (iii) discutir los principales problemas encontrados en las fases de transcripción y alineación sonido/escritura, anotación de los fenómenos de la oralidad (solapamientos, prosodia, aspiraciones, fonética sintáctica, estilo directo, etc.) y en la fase de trasferencia a plataformas de consulta y extracción de datos.

Referencias citadas

Briz, Antonio et alii (1995): *La conversación coloquial. Materiales para su estudio*. Anejo XVI de la revista Quaderns de Filologia.

Briz, Antonio y Val.Es.Co. 2002. *Corpus de conversaciones coloquiales*. *Oralia.* Madrid: Arco.

Cabedo, Adrian y Salvador Pons, en línea (eds.). Corpus Val.Es.Co 2.0. http://www.valesco.es.

Atenuación, ilocución e imagen

Carolina Figueras Bates
Universitat de Barcelona

M. Amparo Soler Bonafont
Universitat de València

La noción de atenuación constituye un tema controvertido en la bibliografía. En tanto que algunos autores la definen como una estrategia para reducir y para compensar los efectos sociales no deseados que el discurso puede provocar en la audiencia (cfr. Fraser 1980; Schneider 2013), otros mantienen en su definición una nítida diferencia entre los aspectos sociales y lingüísticos de la comunicación (Briz 2012; Cestero 2015), a través de la limitación del papel de la imagen al recurso de la atenuación.

El objetivo del presente trabajo es desarrollar un análisis contrastivo de los recursos de atenuación, y su relación con el tipo de acto de habla ejecutado, en dos géneros discursivos en español (conversaciones coloquiales e interacciones en un foro de Internet), con el fin de evaluar la incidencia de la imagen en la manifestación de esta estrategia pragmática. Ambos incluyen la dimensión interpersonal de la comunicación, son dialógicos, no formales, pero se distinguen en cuanto al carácter no mediado y privado de la conversación coloquial, frente al mediado y público de las interacciones del foro. Para realizar este estudio, se han seleccionado dos corpus de 15.000 palabras cada uno, correspondientes a cada uno de los dos géneros escogidos. Los recursos de atenuación empleados en cada subcorpus se han identificando y codificado de acuerdo con la ficha metodológica propuesta en Albelda et al. (2014). A continuación se identificaron los actos de habla principales en cada género, y se examinó la función de las tácticas de atenuación para la articulación ilocutiva del discurso en cada caso a través de la noción de imagen (cfr. Albelda 2016, en prensa).

Bibliografía

Albelda, A. (en prensa): "La variación genérico-discursiva de la atenuación como resultado de la variación de la imagen". *Spanish in Context*.

Albelda, A. (2016): "Sobre la incidencia de la imagen en la atenuación pragmática". *Revista Internacional de Lingüística Iberoamericana*, 27, 1, pp. 19-32.

Albelda et al. (2014): "Ficha metodológica para el análisis pragmático de la atenuación en corpus discursivos del español. (ES.POR.ATENUACIÓN)". *Oralia*, 17, pp. 7-62.

Briz, A. (2012): "La (no)atenuación y la (des)cortesía, lo lingüístico y lo social: ¿son pareja?". En Escamilla, J. y G. Henry (eds.): *Miradas multidisciplinares a los fenómenos de cortesía y descortesía en el mundo hispánico*. Barranquilla/Estocolmo: Universidad de Estocolmo/Universidad del Atlántico/CADIS/Programa EDICE, pp. 33-75.

Cestero, A. M. (2015): "La atenuación lingüística en el habla de Madrid: un fenómeno sociopragmático variable". En Ana M. Cestero, Isabel Molina y Florentino Paredes (eds.): *Patrones sociolingüísticos de Madrid*, Bern, Peter Lang, pp. 365-412.

Fraser, B. (1980): "Conversational mitigation". *Journal of Pragmatics*, 4, pp. 341-350.

Schneider, S. (2013): "La atenuación gramatical y léxica". *Oralia*, 16, pp. 335-352.

Processing native and non-native speech corpus data: a phonological illustration with the French schwa

Romain Isely[1], Isabelle Racine[1], Sylvain Detey[2] & Julien Eychenne[3]

[1]ELCF, Université de Genève, 5, rue de Candolle, 1211 Genève 4, Switzerland, [2]SILS, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku, 169-8050 Tokyo, Japan, [3]Hankuk University of Foreign Studies, Mohyeon, Yongin, 17035 Gyeonggi, South Korea

Abstract

Spoken corpora data processing has greatly benefited from the technological developments of natural speech processing tools over the past decades, both for native and non-native corpora (Kawaguchi, Zaima & Takagaki 2006, Diaz-Negrillo, Ballier & Thompson 2013). In the field of corpus phonology (Durand, Gut & Kristoffersen 2014), such developments have led to sociolinguistically richer descriptions of phonological systems, allowing for data-driven insights into synchronic inter- and intra-speaker variation. In the case of French, the *Phonology of Contemporary French* (PFC) research programme, launched nearly 20 years ago in a Labovian approach (Durand, Laks & Lyche 2002), has become an essential reference for all researchers working on French phonology in the French-speaking world (Gess, Lyche & Meisenburg 2012), with its specific methodology and tools, and with a whole branch dedicated to the study of non-native speakers, in the *InterPhonology of Contemporary French* (IPFC) project (Detey, Racine, Kawaguchi & Zay 2016). In tune with the increased focus set on the sociolinguistic dimension of the speakers' phonological competence by several researchers, the analysis of potential sociolinguistic markers such as *liaison* and *schwa* – two major phenomena in French phonology – have been extensively investigated in PFC (Lyche 2016, Durand & Lyche 2016), but it is only recently that their non-native processing has started to be examined in IPFC, using a similar, and therefore comparable, protocol and methodology. In this presentation, we briefly describe the (I)PFC methodology, with a focus on the perceptual annotation system and the ad hoc software designed to analyze our data. We illustrate our approach with an extended presentation of the alphanumeric code created to probe into the behavior of French schwa produced by non-native speakers in the IPFC-Alemannic corpus of Alemannic-speaking Swiss learners of French. Schwa (also known as "mute e") refers to the French vowel /ə/ – usually written with the letter <e> – that can be either fully phonetically realized or not realized at all (e.g. the word "semaine", i.e. 'week', can be pronounced [səmɛn] or [smɛn]). The ability to successfully manage this alternation between the production and the non-production of schwa is a good indicator of the acquisition of sociolinguistic competence by French L2 learners (Paternostro, Didelot & Racine 2017; Isely, Racine, Detey, Andreassen & Eychenne 2018). In order to get an accurate idea of how the schwa is produced by learners of French, using a speech corpus with the IPFC methodology can be insightful, since it allows us not only to

compare it with native speakers' production (in the PFC native corpus) but also, thanks to the scope of its coding system, to account for the potential influence of several factors at stake. Indeed, the alphanumeric annotation system designed for schwa encodes its word position, its preceding and following phonological contexts, its realization and its quality. The preliminary results of our study show an effect for two variables, the task used (reading vs conversation) and the type of words the schwa appears in (monosyllabic words vs initial or internal positions in polysyllabic words). Besides, when we examine our metadata and distinguish between the learners with and without an immersive learning experience in a French-speaking area, we find an effect of the study-abroad variable when considering its interaction with the other two variables. Finally, a comparison of our non-native results with the productions of the native speakers was also carried out. This methodology has been applied to other surveys within the IPFC project (Racine & Detey 2015), but has also inspired similar projects for English (Kamiyama, Lacoste & Herry-Bénit 2016) and Spanish (Carranza, Cucchiarini, Burgos & Strik 2014; Pustka, Gabriel, Meisenburg, Burkard & Dziallas 2018), and it might be of interest to other researchers in the community of spoken corpora analysts.

References

Carranza, M., Cucchiarini, C., Burgos, P. & Strik, H. (2014). Non-native speech corpora for the development of computer assisted pronunciation training systems, *Proceedings of Edulearn 2014, Barcelona*, 3624–3633.

Detey, S., Racine, I., Kawaguchi, Y. & Zay, F. (2016). Variation among non-native speakers: the InterPhonology of Contemporary French. In S. Detey, J. Durand, B. Laks et C. Lyche (eds), *Varieties of Spoken French.* Oxford : Oxford University Press, 491-502.

Diaz-Negrillo, A., Ballier, N. & Thompson, P. (eds) (2013*). Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam/Philadephia: John Benjamins.

Durand, J., Gut, U. & Kristoffersen, G. (eds) (2014). *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press.

Durand, J., Laks, B. & Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In C. Pusch & W. Raible (eds), *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language*. Tübingen: Gunter Narr Verlag, 93-106.

Durand, J. & Lyche, C. (2016). Approaching variation in PFC: the liaison level. In: S. Detey, J. Durand, B. Laks & C. Lyche (eds), *Varieties of Spoken French*, Oxford: Oxford University Press, 363-375.

Gess, R., Lyche, C. & Meisenburg, T. (eds) (2012). *Phonological Variation in French: Illustrations from three Continents*. Amsterdam/Philadelphia: John Benjamins.

Isely, R., Racine, I. Detey, S., Andreassen, H. N. & Eychenne, J. (2018). Le rôle de l'immersion dans l'apprentissage du schwa chez les apprenants alémaniques avancés de FLE. *Actes du colloque CMLF2018*, Université de Mons, 9-13 juillet 2018, https://doi.org/10.1051/shsconf/20184607010.

Kamiyama, T., Lacoste, V. & Herry-Bénit, N. (2016). The ICE-IPAC project: Testing the protocol on Norwegian and French learners of English, *New Sounds 2016: 8th International Conference on Second Language Speech*, Jun 2016, Århus, Denmark.

Kawaguchi, Y., Zaima, S. & Takagaki, T. (eds) (2006). *Spoken Language Corpus and Linguistic Informatics*. Amsterdam/Philadephia: John Benjamins.

Lyche, C. (2016). Approaching variation in PFC: the schwa level. In: S. Detey, J. Durand, B. Laks & C. Lyche (eds), *Varieties of Spoken French*, Oxford: Oxford University Press, 352-362.

Paternostro, R., Didelot, M. & Racine, I. (2017). Quelques traits stylistiques chez les apprenants italophones de FLE, *Repères DoRiF*, DoRiF Università, Rome, http://www.dorif.it/ezine/show_issue.php?iss_id=23.

Pustka, E., Gabriel, C., Meisenburg, T., Burkard, M. & Dziallas, K. (2018). (Inter-)Fonología del Español Contemporáneo/(I)FEC: metodología de un programa de investigación para la fonología de corpus, *Loquens, 5.1* (DOI: 10.3989/loquens.2018.046. Published online: 12.06.2018), Pustkaetal2018_Loquens.

Racine, I. & Detey, S. (eds) (2015). L'apprentissage de la liaison en français par des locuteurs non natifs : éclairage des corpus oraux. *Bulletin VALS-ASLA* 102.

The ESLORA Corpus of Spoken Spanish: Design, Compilation and Search Engine

Victoria Vázquez Rozas
Universidade de Santiago de Compostela
Mario Barcala
NLPgo Technologies S.L.

ESLORA is a corpus of Spanish made up of semi-directed interviews and spontaneous conversations recorded in Galicia between 2007 and 2015. The initial design had a two-fold objective: to register the use of a variety of Spanish which to date has been scarcely documented and to study the resulting effects of using different techniques for eliciting speech on the registered samples. Every step of the construction of the corpus has compelled us to deal with unforeseen difficulties, which in turn have made us reflect on a range of theoretical and practical aspects involved in the compilation of spoken corpora.

After presenting the main features of the ESLORA corpus, we will discuss some critical questions arising from the coding of "non-standard" varieties in a bilingual context, as is the case in Galicia. Next, we will explain the processing stages that all the documents of the corpus must follow before being included in the search engine and we will also describe its current search capabilities. Finally, we will run some examples to show the usefulness of the search engine (http://eslora.usc.es/).

Resums dels pòsters

Noelia de la Torre Martínez / Universitat de València
nodelat2@alumni.uv.es

L'ATENUACIÓ EN EL CATALÀ I L'ESPANYOL DE VALÈNCIA: ANÀLISI CONTRASTIVA

L'objectiu d'aquesta investigació és realitzar un estudi contrastiu de l'atenuació lingüística (Cafii, 2007) en el català i l' espanyol a la mateixa zona geogràfica: València. Considerem l'atenuació com una categoria pragmàtica, és a dir, un mecanisme encarregat de mitigar el missatge per tal de suavitzar les tensions, amenaces a la imatge pròpia i, especialment, a l' aliena (Briz, 1998).

En aquest sentit, és imprescindible per a realitzar l'estudi d'un fenomen pragmàtic disposar d'un corpus discursiu en el qual el context permet reconéixer el seu ús i la seua funció. Per a desenvolupar l'estudi partim de la proposta metodològica de l'anàlisi de l'atenuació del projecte Es.Var.Atenuación (http://esvaratenuacion.es/) en la qual s'analitza no únicament la forma lingüística, sinó també els paràmetres situacionals, enunciatius i sociolingüístics.

En relació al corpus, analitzarem una mostra total de 3.000 paraules de les dues llengües. Per al cas del català partirem d'un corpus propi gravat en 2018 a València i per al cas de l'espanyol utilitzarem, parcialment, un corpus propi i gravacions del corpus Val. Es.Co (http://www.valesco.es/).

Els primers resultats corroboren el contrast de formes i funcions atenuants entre les dues llengües. D'aquesta manera, hem apreciat l'existència de formes d'atenuació emprades en el català que no s'utilitzen en l'espanyol de València i viceversa, a pesar que utilitzen mecanismes lingüístics en comú.

BIBLIOGRAFÍA

ALBELDA MARCO, Marta et al. (2014): «Ficha metodológica para el análisis pragmático de la atenuación en corpus discursivos del español», Oralia: *Análisis del discurso oral*, nº 17, pp.7-62.

BRIZ GÓMEZ, Antonio (1998): *El español coloquial en la conversación: esbozo de pragmagramática*, Barcelona, Ariel.

BRIZ GÓMEZ, Antonio y ALBELDA MARCO, Marta (2013):«Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español y portugués. La base de un proyecto en común (ES. POR. ATENUACIÓN)», Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile, nº 28, pp.288-319.

CAFII, Claudia (2007): *Mitigation*. Oxford: Elsevier

Dialect contact in a Basque valley

Irantzu Epelde (CNRS-IKER UMR 5478)
irantzu.epelde@iker.cnrs.fr

Recent dialectological work on Basque suggests that a series of linguistic changes are occurring among the Basque native speakers from the southern area (Mitxelena 1961, Zuazo 1998, 2000, 2005, 2013, Aurrekoetxea 2004). This poster presents speech data collected in sociolinguistic interviews in the southern Basque valley of Bidasoa-Txingudi, traditionnaly a geographical and cultural unit located in the border with France. The primary goal of the study is to determine whether apparent-time evidence exists for changes in progress and, if so, who is in the vanguard of these changes.

The poster describes a series of dialect-based changes in progress in the valley, based on data collected in sociolinguistic interviews with twenty local Basque L1 speakers. Dialectal variation in elements chosen from different parts of the grammar are examined: two morphosyntactic alternations on auxiliary verbs and three phonological processes. In particular, several claims are made about dialect contact in Bidasoa-Txingudi. Strong apparent-time evidence exists that four out of five of these elements are undergoing change. What the relevant age limits are remains to be investigated, but older speakers tend toward forms characteristic of the traditional dialect of the valley, whereas younger speakers prefer standard forms (Hualde 1991, Hualde & Ortiz de Urbina 2003).

Language variation can mark stable class differences or stable sex differences in communities, but it can also indicate instability and change (O'Shannessy 2011). When it marks change, the primary social correlate is age (Chambers 2002), and the change reveals itself prototypically in a pattern whereby some minor variant in the speech of the oldest generation occurs with greater frequency in the middle generation and with still greater frequency in the youngest generation. If the incoming variant truly represents a linguistic change (Labov 1994, Trudgill 1974), as opposed to an ephemeral innovation as for some slang expressions or an age-graded change, it will be marked by increasing frequency down the age scale, as it occurs with the youngest generation in this community.

Finally, preliminary results suggest three main directions for further research. First, a more thorough understanding is needed of speakers' attitudes toward these varieties in order to understand all processes of change and dialect contact. Second, much more research is needed into the historical context underlying the gendered distribution of dialectal varieties. Finally, an examination of a broader range of linguistic features is needed to gauge the effects of standardization on the local vernacular.

References

Aurrekoetxea, G. 2004. Estandar eta dialektoen arteko bateratze-joerak: ikuspuntu teorikotik begirada bat [Unification tendencies between standard and dialects: A look from a theoretical viewpoint]. *Uztaro* 50, 45-57.

Chambers, J. K. 2002. Patterns of Variation including Change. In *The Handbook of Language Variation and Change*. Oxford: Blackwell. 349-372.

Hualde, J. I. & J. Ortiz de Urbina. 2003. *A Grammar of Basque*. Berlin: Mouton de Gruyter.

Hualde, J. I. 1991. *Basque Phonology*. Routledge: London.

Labov, W. 1994. *Principles of Linguistic Change: Internal Factors*. Oxford: Blackwell.

Mitxelena, K. 1961. *Fonética histórica vasca*. Diputación Provincial de Guipúzcoa: San Sebastián.

O'Shannessy, C. 2011. Language contact and change in endangered languages. In Peter K. Austin & J. Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

Trudgill, P. 1974. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society* 3: 215-246.

Zuazo, K. 1998. Euskalkiak, gaur [Basque dialects, today]. *FLV* 30, 191-233.

Zuazo, K. 2000. *Euskararen sendabelarrak* [The medicinal herbs of Basque]. Irun: Alberdania.

Zuazo, K. 2005. *Euskara batua: ezina ekinez egina* [Standard Basque: The impossible accomplished with effort]. Donostia: Elkar.

Zuazo, K. 2013. *The dialects of Basque*. Reno, Nevada: Center for Basque Studies, University of Nevada.

Construing scientific knowledge and politeness in ten doctoral defenses held in Finnish and in French – an exploratory study of spoken academic discourse

Hanna Jokela and Milla Luodonpää-Manni, University of Turku

The purpose of the presentation is to explore spoken academic discourse from the perspective of doctoral defenses. Although academic discourse in general has been subject to numerous studies in the past few decades, its spoken variety has received relatively little attention (see however Hartwell et al. 2017). On the one hand, this can be explained by the fact that written language has traditionally had a strong position in spreading scientific knowledge. On the other hand, the study of spoken academic discourse faces many practical challenges that are typically related to the collection, manipulation and analysis of spoken corpora. Even though scientific research articles and doctoral dissertations undeniably have central position in spreading scientific knowledge, scientific knowledge spreads also – and often in the first hand – through oral presentations held in the scientific community.

In this study we explore the mechanisms in scientific knowledge construction and the politeness strategies used in public doctoral defenses through the expressions of agreement and disagreement. Public doctoral defenses offer an interesting material for a study dealing with scientific knowledge construction because one of the objectives in public doctoral defenses is to discuss the claims made by the doctoral candidate and to decide which claims may be accepted as scientific knowledge and which may not. At the same time, public doctoral defenses have their own special character as highly formal academic situations, which increases the need for adopting special politeness strategies (e.g. Goffman 1967), especially in relation to the expression of disagreement. Jokela (2012), for example, has shown that person references play an important role in politeness strategies. Finnish and French have both similar and different politeness strategies in formal situations, and both languages express personal reference in various ways (Helasvuo & Johansson 2008).

In the present study we aim to answer the following questions: What is the role of the expressions of agreement and disagreement in the scientific knowledge constructions in the context of public doctoral defenses? What kinds of personal reference constructions are used by the speakers in reference to oneself and to the other participant? Which strategies are used in order to save one's own face and that of the other? To answer these questions, we analyse a spoken corpus of ten doctoral defenses recorded between 2007-2013 in three Finnish universities (University of Turku, University of Helsinki and Åbo Akademi, 20 hours in

total). The corpus consists of recorded doctoral defenses held in two languages, Finnish and French, five defenses each. As methods of analysis we use a combination of linguistic analysis, conversation analysis and concept analysis.

Bibliography

Goffman, Erving.1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. New York: Doubleday.

Hartwell, L., Esperança-Rodier, E. & Tutin, A. 2017. I think we need…: Verbal expressions of opinion in conference presentations in English and in French. *CHIMERA: Romance Corpora and Linguistic Studies* 4(1), 35–60.

Helasvuo, M-L. & Johansson, M. 2008. Construing reference in context: Non-specific refernce forms in Finnish and French discussion groups. In María de los Ángeles Gómez González, J. Lachlan Mackenzeie & Elsa M. González Álvarez (eds.), *Current Trends in Contrastive Linguistics. Functional and cognitive perspectives*, 27–57. Studies in functional and structural linguistics 60. Amsterdam/Philadephia: John Benjamins Publishing Company.

Jokela, H. 2012. Nollapersoonalause suomessa ja virossa: tutkimus kirjoitetun kielen aineistosta [Zero person clause in Finnish and Estonian]. Turku: University of Turku. Available: http://urn.fi/URN:ISBN:978-951-29-4912-0.

Building up a multi-purpose reference corpus of spoken interactions

Julia Kaiser, Evi Schedl & Thomas Schmidt

With the Research and Teaching Corpus of Spoken German (FOLK), the Institute for the German Language (IDS; member of the Leibniz Association) is building up a large corpus of authentic, spontaneous spoken interactions, recorded on audio and/or video. FOLK currently comprises 281 interactions with a length of 230h, 2.2 million tokens and 806 speakers (varying in gender, age, educational background, and regional provenance). It is intended as a reference corpus, to be used in research and teaching, by scholars and students, in different disciplines (or at least different subfields of linguistics), and with diverse qualitative and quantitative methodological approaches. With this claim of "universality" – i.e. with consciously abstaining from tying the corpus to a specific usage scenario – comes a variety of challenges concerning corpus design and stratification, data collection, methods of documentation and annotation, and tools for disseminating and analysing the corpus data (Schmidt 2016; 2017). In our contribution to the workshop, we would like to focus on a selection of these challenges.

A first challenge lies in the task of **corpus design**. As a reference corpus, FOLK must strive to represent its subject domain – spoken interactions – in as broad and differentiated a manner as possible, and it must make the underlying systematics transparent to the user. We use "interaction type" as the primary parameter in corpus stratification, meaning that we view differences in interaction constellations, contexts and content ("situational parameters" according to Biber 1993: 245) as the most important parameters requiring an adequate representation in the corpus. We therefore combine a more global distinction of three interaction domains (private, institutional, public) with finer-grained parameter sets for spheres of social life and activities, to approach the “communicative household” (Luckmann 1988) of the German speaking society. Socio-demographic properties of speakers (see above) are taken into account to balance the corpus data on a second level. Since, however, data acquisition for this type of corpus cannot be planned fully in advance (because it depends to a considerable degree on opportunities of field access), and since there is no widely established classification system for interactions, corpus stratification has to be continually adjusted as the corpus keeps growing. For this purpose, we have developed methods for assessing corpus balance at each stage and strategies for effectively acquiring data from different sources.

A second challenge is **corpus annotation**. For transcription of the audio and video recordings, different research approaches require different levels of detail on different linguistic levels (such as lexis, phonetic, prosody and others). Not only does a multi-purpose

corpus have to compromise between these requirements, it also has to trade off the desirable amount of precision against the project's finite capacities for the time-consuming transcription process. We have decided to use a simplified version of a widely used transcription system (GAT2, Selting et al. 2009) for initial transcription, which is well suited to support conversation analytic and other qualitative oriented approaches to the data. Additional annotation layers for orthographic normalisation, lemmatisation and POS tags are added to the transcription in order to cater for more quantitative oriented approaches such as corpus linguistics. A set of interoperable tools efficiently supports the annotation workflow and automates steps wherever possible.

For completed corpus data, **dissemination** is a third important challenge. Again, user expectations to an optimal presentation of the corpus data will vary considerably depending on the researchers' backgrounds and methodological approaches. We currently make FOLK available via the Database for Spoken German (DGD), an internet platform for oral corpora, which attempts to combine tools for qualitative approaches to the data (browsing and viewing/listening to metadata, transcripts and recordings) with query mechanisms, which are known from corpus linguistics, but have to be adapted to better suit the work with spoken data.

References

Biber, Douglas (1993): Representativeness in Corpus Design. Literary and Linguistic Computing, Vol. 8, No. 4, 243-257.

Luckmann, Thomas (1988): Kommunikative Gattungen im kommunikativen "Haushalt" einer Gesellschaft. In: Gisela Smolka-Koerdt/Peter M. Spangenberg/Dagmar Tillmann-Bartylla (Hg.): Der Ursprung von Literatur. München, 279-288.

Schmidt, Thomas (2016): Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. In: Compilation, transcription, markup and annotation of spoken corpora. In: Kirk, John M./Andersen, Gisle (Hg.): Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], 396-418.

Schmidt, Thomas (2017): Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. In: Kupietz, Marc/Geyken, Alexander (Hg.): Corpus Linguistic Software Tools, Journal for Language Technology and Computational Linguistics (JLCL 31/1), 127-154.

Selting, Margret et al. (2009) Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, 353-402.

Give it me back on our wedding day: On the alternative double object construction on spoken Asian varieties of English.

Juan Lorente Sánchez
University of Málaga

Abstract

The concept dative alternation refers to the construction in which a ditransitive verb takes a subject and two objects referring to a theme and a recipient (Gast 2007:31). In English, this phenomenon offers the possibility of the alternation between a prepositional object construction (PREP), where the recipient is encoded as a prepositional phrase (*Give it to me*), and a double object construction (DOC), where the recipient takes precedence over the theme (*Give me it*). In addition, the double object construction may also appear with an alternative double object variant (altDOC) if non-standard varieties of English are considered (*Give it me*). Given that this ditransitive construction is the least frequent type in standard British and American English and since "few studies have considered language-external (i.e. regional, diachronic or social) factors in determining the dative alternation" (Gerwin 2013:20), the present paper aims to analyze the use of alternative double object construction in some spoken Asian varieties of English. Attention will be paid to the gender, the age and the social class of the informants and other determining factors such as the type of object (pronominal or nominal) along with the semantic and pragmatic environment in which it occurs. Thus, this paper pursues the following objectives: a) a quantitative study of the phenomenon in these varieties of English; b) a variational analysis of these alternative forms in terms of the age, gender and social background on the informants; and c) a classification of these variant forms from a semantic and pragmatic standpoint. The corpus used as source of evidence comes from the *International Corpus of English* (ICE), i.e. Indian English, Hong Kong English and Singapore English.

References

Levin, B. 1993. *English verb classes and alternations.* Chicago: The University of Chicago Press. 45-48

Rappaport-Hovav, M. & Levin, B. 2008. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics.* 129-163

Szmrecsanyi, B., Grafmiller, J., Bresnan, J., Rosenbach, A., Tagliamonte, S., & Todd, S. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2(1): 86. 1-27, DOI: https://doi.org/10.5334/gjgl.310

Gerwin, J. 2013. *Ditransitives in British English Dialects.* Berlin: Walter de Gruyter. 166-199

Gast, V. 2007. *I gave it him* – on the motivation of the 'alternative double object construction' in varieties of British English. In (pp. 31-56) Siewierska, A & Hollmann, W. (eds). *Functions of Language – Special issue: Ditransitivity*.

Nausica Marcos Miguel, marcosn@denison.edu
Claudia Sánchez Gutiérrez, chsanchez@ucdavis.edu

A Spanish Second Language Classroom Corpus: Discussing its construction and shareability

Compared to corpora available on *English as a Second Language* (ESL) learners and classroom, there are limited resources for researchers on *Spanish as Second Language* (SSL). The classroom is a special place of interest for researchers since it is one of the main spaces where learning happens and where input surrounds the learners in formal instructional settings. However, there is currently no available SSL classroom corpus.

This presentation will discuss the beginning stages of the compilation of a *SSL Classroom Corpus*. Given the variety of classroom contexts of SSL, this corpus focuses on Spanish taught at US higher-education institutions in the first two years of instruction, i.e., elementary and intermediate language courses.

Data has been collected from three institutional settings. Two PhD-granting universities (A and B), where Teaching Assistants pursuing a PhD comprise most of the teaching force in foreign language courses, and a liberal arts college (C), where more experienced faculty tend to teach language courses. Data collection in B and C is still ongoing. So far, the data collected comprises:

1. 16 classroom sessions of three sections of a multi-section intermediate language course taught each section by a different instructor (three teachers in total) (A);
2. between 30 and 50 class session of three sections of a multi-section elementary Spanish course taught each section by a different instructor (three teachers in total) (B);
3. 40 class sessions from three courses, a beginner and two intermediate language course, taught by the same instructor (C).

Whereas more data will be collected in the following semesters in elementary and intermediate language courses in B and C, the already collected data is currently been transcribed. The target is to collect data from 10 instructors teaching beginning and intermediate levels so that comparisons can be drawn among levels and instructors.

Overall, this corpus serves to analyze the input provided in those classrooms by the instructors. Users of this corpus will be researchers interested in SSL learning and teaching, as well as those interested in teacher development. It will be possible to perform longitudinal and cross-sectional analysis of instructors. For example, this corpus can help to answer questions related to:

1. Vocabulary frequencies in teacher talk
2. The role of teaching materials in the classroom
3. Analysis of morphosyntax and pragmatics in teacher talk
4. Variation in teacher talk

In the presentation, there will be time to discuss some of the challenges of this project. First, due to IRB constraints, the data needs to be audio recorded and cannot be video recorded. Second, whereas the teacher can be always identified, this is not the case

of the learners who can not be followed beyond one specific class. Thus, the corpus only allows for a thorough analysis of the teacher talk.

Part of the discussion will also deal with the best approach for sharing classroom corpora. Similar corpus of ESL classrooms are restricted to researchers (e.g., Reder, Harris, & Setzler, 2003; Trofimovich, Collins, Cardoso, White & Horst, 2012), whereas the transcriptions are publicly available for others (e.g., Jäkel, 2010). For this corpus, sharing anonymized transcriptions without recording seems to be the most suitable way to preserve the confidentiality of the participants. Providing plain text files to other researchers will also help them to develop a system to answer their own research questions.

In brief, this presentation will focus on the construction of a SLL classroom corpus and on the challenges this project presents to build a representative as well as shareable corpus.

EL VALOR ATENUANTE DEL DISCURSO DIRECTO DE PENSAMIENTO EN EL ESPAÑOL DE AMÉRICA

DAVID NAVARRO CIURANA

El discurso directo de pensamiento (DD-p) supone una reconstrucción oral del lenguaje interno (Benavent, 2015) e implica, a través de un fingimiento deliberado de mímesis, la consumación de un retrato completo, aunque filtrado, del emisor, como podemos ver en los siguientes ejemplos: «dije *¿pa qué me va a servir?* y la boté»; «en un momento la Ju-bueno yo ehtaba conversando con Juancito adelan- un poco máh adelante y elloh ehtaban atráh hablando bueno yo digo *la Laura se va a copaar y va a ehtar hablando acá hahta las ocho de la noche*».

El objetivo del póster es estudiar el DD-p en un corpus de conversaciones coloquiales del español de América (Ameresco), basándose únicamente en las conversaciones grabadas en La Habana (Cuba) e Iquique (Chile). Tras el análisis de una serie de factores como la función discursiva de las ocurrencias de DD-p, su atribución o su elemento introductor, podemos esbozar la hipótesis de partida de que muchos de estos usos responden a un valor atenuante y de mejora de imagen del locutor. Se trata de una función que no es de naturaleza correctiva, sino que responde a una estrategia de aproximación, expresión y reafirmación del yo ante el receptor (Hernández Flores, 2004), que incrementa el vínculo de familiaridad entre los hablantes, dado que supone compartir aquello que pertenece a la esfera más íntima. Además, se estudia una posible variación geográfica de esta función en las distintas ciudades americanas que recoge el corpus, atendiendo al mayor o menor uso del DD-p así como a la función discursiva que este posea en cada ocurrencia.

PALABRAS CLAVE

Lingüística de corpus

Discurso directo

Atenuación

Español de América

BIBLIOGRAFÍA

Albelda Marco, M. (2012). La atenuación lingüística como fenómeno variable. *Oralia*, *15*, 77-124.

Benavent Payá, E. (2003). ¿ Por qué contamos nuestras historias cotidianas en estilo directo?. *Foro hispánico: revista hispánica de Flandes y Holanda*, (23), 11-20.

Díaz, Y., & Labarca, M. (2010). Las formas digo, vamos a decir, dicen, como marcadores discursivos (intensificadores, atenuadores y justificadores) en el habla de Mérida, Venezuela. *Lengua y Habla*, *14*(1), 12-24.

Flores, N. H. (2004). La cortesía como búsqueda del equilibrio de la imagen social. In *Pragmática sociocultural: estudios sobre el discurso de cortesía en español* (pp. 95-108). Ariel.

Flores Treviño, M. E. (en línea). Corpus de conversaciones Ameresco-Monterrey, en Albelda y Estellés (coords.). Corpus Ameresco. www.esvaratenuacion.es

A collection of non-standard spoken Russian corpora: approaches, tools and research

Ruprecht von Waldenfels, FSU Jena
Anastasia Panova, Linguistic Convergence Laboratory NRU HSE

The paper presents several corpora of the varieties of Russian spoken in different regions. The corpora were created in 2013-2018 and use a similar methodology in data collection, annotation and presentation. These corpora make it possible to study phonetic and morphosyntactic characteristics of these varieties both qualitatively as well as quantitatively. The general approach as well as the computational tools are shared with partners in Poland and Germany in the Spoken-Slavic network.

The corpora consist of interviews recorded during field trips of affiliated groups in different regions of Russia in 2007-2018. The overall approach of corpus building is to transcribe the interviews using ELAN and Praat in standard orthography without phonetic detail, and to rely on automated tools for further processing (see Waldenfels, Daniel, Dobrushina 2014). Currently, the texts are then processed automatically to add lemmatization and morphological tagging. Work is underway to automatically align the transcription and the audio on a phoneme level; the phonetic level should be available for query in 2019. The corpora are made online available with SpoCo (Waldenfels, Woźniak 2017), an interface which allows complex queries and gives back both transcription and links to audio segments.

One part of collection consists of dialects corpora. The latest corpus of this type represents the idiolects of ten indigenous inhabitants of the village Rogovatoye (Rogovatka) in the Starooskolsky region of the Belgorod Oblast (south-western part of Russia). The Rogovatka dialect has been studied by dialectologists in 2012-2015 (Bukrinskaja et al. 2014), and this new corpus allows to make further, quantitatively oriented investigations of this variety. At present, this corpus comprises 10 hours and 100,047 tokens. A further dialectal corpus is currently currently being prepared with texts collected in 2017 in the Bryansk Oblast (western Russia); that corpus comprises 12 hours and 80,314 tokens. The third dialectal corpus, the Ustja River Basin Corpus, created in 2013-2017, is the largest and currently comprises 97 hours and 767,149 tokens of the speech of dozens of speakers.

The research associated with this latter corpus is specifically concerned with the dynamics of dialect change and language shift from the dialect to the standard variety in a number of apparent time studies based on different variables. A sample graph concerning the use of the variable ‘lack of prothetic n with pronouns” is shown in figure 1. Tools and procedures are developed during this line of research that is later put into use in respect to the others.

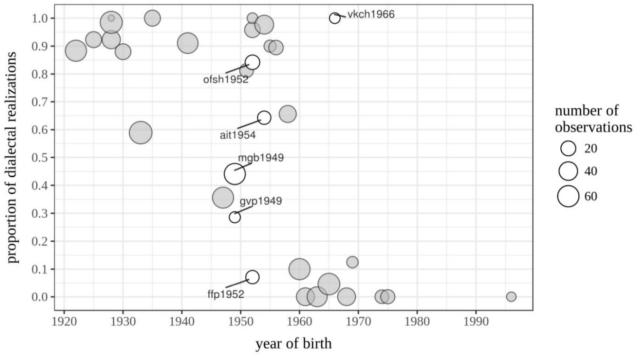

Figure 1

Another type of corpus represented in the collection are L2 varieties of Russian that are spoken as lingua franca in different regions. Currently one such corpus is ready: The texts in the corpus of Russian spoken in Daghestan (eastern Caucasus) were produced by people whose native languages belong to Nakh-Daghestanian or Turkic families, but all of them speak Russian and use it as a lingua franca. Such type of corpora gives a rich material for investigating interference with languages typologically and genetically distant from Russian. Some properties of Russian spoken in Daghestan are discussed in papers (Daniel et al. 2010; Daniel, Dobrushina 2013). A second such corpus of Russian spoken in Chuvashia which represents speech of Russian-Chuvash bilinguals (Chuvash is a Turkic language) is currently under development.

These corpora make it possible to study individual varieties in more depth than previously possible, especially combining the intersection sociolinguistic and systemic issues of non-standard Russian varieties. Furthermore, a systematic comparative study of linguistic features of regional varieties of Russian becomes feasible, as more corpora are added.

References

Bukrinskaja I. A., Djachenko S. V., Karmakova O. E., Ter-Avanesova A. V. 2014. Otchety o dialectologicheskikh expedicijakh Instituta russkogo jazyka im. V.V. Vinogradova RAN v 2013 g. [Reports on dialectological expeditions of the V.V. Vinogradov Russian Language Institute of the RAS in 2013]. *Russkij jazyk v nauchnom osvesh'enii* 2 (28), 262–309.

Daniel M. A., Dobrushina N. R. 2013. Russkij jazyk v Dagestane: problemy jazykovoj interferencii [A corpus of Russian as L2: the case of Daghestan]. In: *Komp'juternaja lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog»*. Vyp. 12. [Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 12]. Moscow: Izd-vo RGGU, 186-211.

Daniel M., Knyazev S. V., Dobrushina N. 2010. Highlander's Russian: Case Study in Bilingualism and Language Interference in Central Daghestan. In: *Instrumentarium of Linguistics: Sociolinguistic Approach to Non-Standard Russian*. Helsinki: Slavica Helsingiensia, 65-93.

Waldenfels R. von, Woźniak M. 2016. SpoCo - a simple and adaptable web interface for dialect corpora. *Journal for Language Technology and Computational Linguistics* 31, 155-170.

Waldenfels R. von, Daniel M. A., Dobrushina N. R. 2014. Why Standard Orthography? Building the Ustya River Basin Corpus, an online corpus of a Russian dialect. In: *Komp'juternaja lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog». Vyp. 13*. [Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 13]. Moscow: Izd-vo RGGU, 720-728.