Mini-Corpus del español para DB-IPIC

Carlota Nicolás Martínez°, Marina Lombán Somacarrera*[1]
°Università degli Studi di Firenze, *Universidad Complutense de Madrid

The paper presents the IPIC extension to a Spanish mini-corpus, conceived and tagged according to the L-AcT criteria and comparable with the Italian and the Brazilian Portuguese mini corpora already stored in the Data Base. IPIC was created by the LABLITA team for grounding the cross-linguistic comparison of Information Structure in spontaneous speech and can be applied to corpora, collected according to a similar corpus design and size (approx. 5000 reference units). The reference units of speech analysis (utterance, stanza) are identifying through terminal prosodic breaks and aligned to the corresponding sound. In turn, reference units are segmented through non-terminal prosodic breaks in information units that are tagged according to the L-AcT repertory of information functions. The paper shows typical Spanish examples of each information unit type, detailing in the case also their specific terminology. The corpus informationally tagged allows to carry out a solid and deep analysis also at the grammatical and lexical levels of spoken Spanish.

Keywords: Corpus linguistics, Spoken Spanish, Information structure, Prosody, Lexicon

1. Introducción

El objetivo de este artículo es describir el contenido y la utilización del Mini-corpus del español oral, que está recogido en la base de datos DB-IPIC junto a los Mini-corpus del italiano (Panunzi & Gregori 2012) y del portugués brasileño (Panunzi & Mittmann, 2014; Raso & Mello 2010, 2012; Mittmann & Raso 2012). Estos tres corpus están estructurados del mismo modo para realizar estudios comparativos entre estas tres lenguas orales espontáneas. Cada uno de estos mini-corpus está etiquetado separando las unidades prosódicas y atribuyendo a cada una de ellas la unidad informativa correspondiente. Estas etiquetas y el análisis de PoS de todos los textos transcritos y alineados con la voz completan un corpus interrogable para el estudio paradigmático y sintagmático de las unidades

[1] Los capítulos 1 y 2 han sido redactados por Lombán; el 3 y el 4 por Nicolás.

CHIMERA. Romance Corpora and Linguistic Studies 5.2 (2018), 95-113. ISSN 2386-2629

informativas IUs del habla y de sus componentes gramaticales y léxicos. Los resultados de las búsquedas se presentan mostrando los enunciados –transcritos y etiquetados- junto al audio correspondiente.

2. Marco teórico

En el análisis prosódico se observa que el habla se estructura en secuencias prosódicas terminadas TS, que pueden estar compuestas por secuencias prosódicas no terminales denominadas unidades tonales TU.

Tras este análisis prosódico de las secuencias terminadas TS -que se realiza durante el proceso de transcripción y etiquetado- se ha observado que, dependiendo de sus características, estas pueden ser Enunciados (ej. 1 y 2) y Estancias (ej. 3). Estas son las dos únicas unidades articulatorias referenciales del habla según la *Language into Act Theory L-AcT* (Cresti 2000; Cresti 2018).

Las TS pueden estar compuestas por secuencias prosódicas no terminales y no independientes llamadas unidades prosódicas TU. A cada una de estas unidades prosódicas le corresponde una unidad de la articulación de la información. Las TU y las TS son componentes esenciales en los tres minicorpus de DB-IPIC (Panunzi & Mittmann 2014: 133).

Es obligado describir la teoría marco en la que nos hallamos. En ella se afirma que cada secuencia lingüística que está marcada por una linde prosódica terminada es completa por lo que respecta a su estructura de la información y equivale a una unidad de referencia del habla espontánea (Panunzi & Mittmann 2014: 130). Las unidades de referencia del habla espontánea son el Enunciado y la Estancia.

En este marco teórico es esencial tener en cuenta que cada Enunciado corresponde a un Acto de Habla; para la teoría *Language into Act Theory* (Cresti 2000; Cresti & Moneglia 2010) un Enunciado es la unidad de referencia para el estudio del habla espontánea y se entiende siguiendo la tradición pragmática de Austin como "the minimal linguistic entity such that can be pragmatically interpreted" (Austin 1962) y sucesiva (Biber *et al*.1999). Son, por lo tanto, las unidades mínimas necesarias, independientes y concluidas. El Enunciado tiene propiedades prosódicas y, por tanto, "The criterion for the segmentation of the speech flow into utterances is prosodic. Each prosodic break qualified as terminal defines the utterances limits in the speech flow" (Cresti & Moneglia 2006: 92).
La Estancia se diferencia del Enunciado en que no es un modelo prosódico con un programa unitario, sino un proceso de anexión de información que sigue el fluir del pensamiento del hablante.

2.1 Unidades referenciales

Los Enunciados corresponderán a secuencias prosódicas que han sido terminadas, que tienen un modelo (*pattern*) prosódico, que realiza un acto de habla y que sigue un programa unitario.

El núcleo del Enunciado es el Comment (1) que es la unidad informativa que realiza la fuerza ilocutiva del enunciado y es necesaria y suficiente para la realización del mismo. Cuando el Enunciado está formado por una sola unidad de Comment se considera Enunciado Simple.

(1) el intento es ese //^{COM} [epubmn03-195]

Un unidad informativa Comment múltiple se realiza cuando el Enunciado está compuesto por una cadena de dos o más unidades de Comment que forman un único modelo ilocutivo en el que se realiza la fuerza ilocutiva del Enunciado; a nivel ilocutivo produce un efecto retórico convencional, como por ejemplo, un refuerzo, una comparación, una lista etc. .

(2) quédese usted ahí fuera /^{CMM} por si surge algo //^{CMM} [epubdl02b-193]

La Estancia es la otra unidad de referencia junto al Enunciado, y al igual que este es una secuencia prosódica terminada TS. Como ya se ha dicho se diferencia del Enunciado en que no es un modelo prosódico que sigue un programa unitario, sino que consiste en un proceso de anexión de información que sigue el fluir del pensamiento del hablante. Por tanto, la Estancia está formada por una secuencia de unidades prosódicas que no sigue un modelo informativo y que tiene un débil valor ilocutivo. Estas unidades prosódicas corresponden a la unidad informativa llamado Comment Ligado. La Estancia[2] corresponde a una actividad lingüística cuya primera intención es la producción de un texto oral no planeado previamente.

(3) el caldo se hace /^{TOP} con pimentón /^{COB} el ajo refrito /^{COB} y se echa al [/1] al puchero /^{COB} a la cazuela /^{COB} donde sea //^{COM} [efamdl02-8]

[2] En el etiquetado de las Estancia los *Comment ligados (Bound Comment)* son marcados con la etiqueta COB, salvo el último que forma la secuencia que se etiqueta con COM. La Estancia además de los COB puede estar compuesta por otras unidades informativas que se verán más adelante.

2.2 Unidades de la articulación de la información

En el etiquetado se atribuye a cada unidad prosódica la etiqueta de la unidad informativa que realiza. Para hacer esto, en este Mini-corpus se han escuchado una a una las unidades prosódicas transcritas y alineadas, y se han observado atentamente las curvas melódicas. Antes se ha estudiado el proceso de análisis que ha llevado a la consolidación de la teoría de la articulación de la información. Esta teoría se ha fraguado gracias al estudio directo de miles de unidades prosódicas en italiano y en portugués hasta llegar a la delimitación de todas las unidades actuales que son las únicas que se pueden considerar que componentes de la articulación de la información.

2.2.1 Unidades informativas textuales nucleares

Las unidades informativas textuales (Tabla 1 y 2) realizan el contenido semántico y proposicional del Enunciado o de la Estancia. Las textuales nucleares son indispensables para que haya un Enunciado o una Estancia. Presentan en el Mini-corpus las siguientes etiquetas:

Tabla 1.

Unidades informativas textuales nucleares

Nombre	**Etiq.**	**Defininición**
Comment (Comment)	COM	Es la unidad informativa que realiza la fuerza ilocutiva del Enunciado y es la unidad necesaria y suficiente para la realización de un Enunciado.
Multiple Comment (Comment Múltiple)	CMM	Es una cadena de dos o más unidades de Comment que forman un único modelo ilocutivo en el que se realiza la fuerza ilocutiva del Enunciado. El modelo ilocutivo realiza solo un Enunciado, que a nivel ilocutivo produce un efecto retórico convencional. Como por ejemplo un refuerzo, una comparación, una lista etc.
Bound Comments (Comment Ligado)	COB	Es una secuencia de Comment que no componen un Enunciado sino una Estancia (Stanza). Una Estancia se produce mediante la adjunción progresiva de unidades de Comment que coinciden con el fluir del pensamiento del hablante. La fuerza ilocutiva de los Comment Ligado (Bound Comment) es débil y homogénea y suele ser asertiva o expresiva.

Algunos ejemplos representativos de estas unidades son de COM:

(4) *LUP: o varios millones de euros // COM
[epubcv01-4]

(5) *MRO: no me lo puedo creer // COM
*BEG: son dos parejas // COM
*BEG: muy bien //COM
[efamcv08-82-84]

Para las unidades de CMM mostramos dos casos:

(6) *DIE: lo ves /CMM o no ? CMM [efamcv07-105]

(7) *BEG: no / CMM chica no // CMM [efamcv08-119]

Un ejemplo de COB, en un diálogo público:

(8) *VIG: te irías aquí /COB a este icono de la carpeta //COM [epubdl07a-221]

2.2.2 Unidades informativas textuales no nucleares

Las unidades informativas textuales, no nucleares, (Tabla 2) son unidades que participan en la composición del contexto semántico del enunciado integrando el Comment.

Tabla 2.

Unidades informativas textuales no nucleares

Nombre	**Etiq.**	**Definiición**
Topic (Tçopic)	TOP	Representa el campo pragmático de aplicación de la actividad lingüística realizada por el Comment. Ofrece la posibilidad de que el Comment se interprete sin necesidad del contexto.
*List of Topic*s (Lista de Tópic)	TPL	Es una cadena de dos o más Topic cuya función es la de un solo Topic; forman parte del mismo campo semántico.
Appendix of Comment (Apéndice de Comment)	APC	Completa el texto del Comment, ya sea mediante un añadido, un proceso de

		repetición estratégica o una información más específica dirigida al interlocutor, siempre con la clara intención de conseguir su aprobación.
Appendix of Topic (Apéndice de Tópic)	APT	Completa el texto del Topic al añadir información complementaria, correcciones y, en raras ocasiones, repeticiones.
Parenthesis (Paréntesis)	PAR	Es una inserción metalingüística relacionada con el contenido del Enunciado, se puede referir a lo dicho anterior o sucesivamente.
Locutive Introducer (Introductor Locutivo)	INT	Indica que el conjunto de las unidades informativas a las que precede, incluido el Comment, tienen un único punto de vista que a menudo se diferencia del punto de vista del Enunciado. Las unidades informativas a las que precede pueden corresponder a un discurso indirecto, a un pensamiento en voz alta, a una lista, a una narración o a una ejemplificación.

9 (9) *LOG: si surgiera algo hoy /TOP que llamasen /APT yo te llamo //COM [epubdl02a -67]

10 (10) *ALV: de los políticos de la República hhh /TPL de antes de la República hhh /TPL de después [/2] de después de la Guerra hhh /TPL eran [/1] eran abogados que se dedicaban a la política //COM [efammn07a-14]

11 (11) *LUP: si gustan /TOP es porque la gente también se reconoce /COM en eso //APC [epubcv01-37]

12 (12) *VIG: pues imagínate /INT que vienen seis personas /TOP y te pagan /CMM y no te dicen nada /CMM no te avisan //CMM [epubdl07a-206]

13 (13) *LOU: entonces /DCT pues al cabo del año /TOP que se puso ya robusta y grande /PAR se hicieron las fotos oficiales de casados //COM [efamdl07-8]

2.2.3 Unidades informativas dialógicas

La Tabla 3 contiene las unidades informativas dialógicas que también pueden estar presentes en los Enunciados Compuestos o en las Estancias, cuyo papel está relacionado con el desarrollo de la interacción, es decir, con la buena realización del intercambio comunicativo. No participan en el contenido proposicional del Enunciado, ya que acometen tareas como mantener el canal de comunicación abierto, expresar cohesión social en relación con el interlocutor o tomar y mantener el turno de palabra.

Tabla 3.
Unidades dialógicas

Nombre	**Etiq.**	**Definición**
Incipit (Incipit)	INP	Abre el canal comunicativo para iniciar un turno o para crear un Enunciado en contraste afectivo con respecto al anterior.
Conative (Conativo)	CNT	Empuja al destinatario a participar de forma adecuada en el intercambio comunicativo, induciéndole a realizar, detener o evitar una acción relacionada con la comunicación.
Phatic (Fático)	PHA	Controla que el canal comunicativo se mantenga abierto y que el intercambio y su recepción se mantengan. Se puede utilizar para mejorar el desarrollo o la conclusión de los Enunciados.
Allocutive (Alocutivo)	ALL	Identifica al destinatario del Enunciado para conseguir su atención, pero al mismo tiempo establece una conexión personal con él.
Expressive (Expresivo)	EXP	Funciona como un soporte enfático del intercambio que se ocupa de la cohesión social entre los participantes del evento comunicativo.
Discourse Connector (Conector Discursivo)	DCT	Relaciona diferentes partes del discurso. Señala al receptor que el discurso continúa y que, por tanto, lo que dirá a continuación tiene relación con lo anterior.

Como vemos en (13) estas unidades informativas (concretamente INP en el ejemplo) tienen necesariamente que estar en el mismo enunciado que una de las unidades nucleares (COM) puesto que no son independientes, y pueden o no estar presentes unidades textuales no nucleares (TOP y PAR).

(14) *MAA: a ver /INP si a lo tonto /TOP ya en realidad /PAR tengo bastante material //COM [efamcv07-195]

Las unidades de Incipit se hallan en posición inicial de los enunciados, aunque a veces no sea inicial absoluta. Lo mismo sucede con algunas otras unidades, como por ejemplo PHA, que se presenta en general al principio o al final de los enunciados.

(15) *ALB: vamos /PHA me ha parecido que era él //COM [epubcv05-35]

Mientras los Conectores Discursivos no se encuentran nunca en posición final:

(16) *PAC: porque /DCT queda todavía /SCA el pringue ese de la obra //COM [efamdl04-33]

Los Conativos no son unidades muy frecuentes.

(17) *CAR: mira mira /CNT qué bien has &em [/1] desempaquetado tú /COM lo tuyo //APC [efamcv03-91]

Tampoco lo son los Alocutivos y los Expresivos:

(18) *LOG: pero es que no hay nada /COM Elena //ALL [epubdl02b 155]

(19) *CAR: hhh por dios /EXP nos lleva la &ve [/1] < &vie [/1] viento > //COM [efamcv04-28]

2.2.4 Unidades no informativas

En la Tabla 4 se presentan las etiquetas para marcar partes del discurso que no tienen ningún valor informativo. Destaca la Escansión pues se ha observado que una sola unidad informativa puede estar dividida en dos o más unidades prosódicas; esto se produce porque la unidad informativa puede tener un pesado contenido locutivo, que necesite ser dividido en dos unidades prosódicas, o porque el hablante ha dividido la unidad informativa en dos unidades prosódicas por razones enfáticas. Existen además otras unidades realizadas debido a problemas de fluidez en el habla como las interrupciones o tomas de aire.

Tabla 4.

Unidades no informativas

Nombre	Etiq.	Definición
Scanning (Escansión)	SCA	Unidad prosódica que no tiene función informativa en sí misma y cuyo valor ilocutivo es el de la unidad informativa de la que hace parte.
Interrupted (Interrupción)	EMP	Es una unidad que ha sido interrumpida y que no puede ser evaluada.
Time Taking (Toma de tiempo)	TMT	Es una unidad de toma de tiempo, usada por necesidades de programación y/o para mantener el turno de palabra.
Unclassifiable (Inclasificable)	UNC	Una unidad inclasificable debido a los insuficientes datos acústicos.

De estas unidades la más frecuente es la de SCA:

20 (20) *PAC: la nevera /TOP habría que /SCA limpiarla por dentro //COM [efamdl04-10]

21 (21) *SIN: sí que ha cambiado un poco /SCA su correlación de fuerza /COB porque requiere /SCA < de otro partido > /COB para aprobar las cosas /COB como es el mismo Ciudadanos //COM [epubmn03-102]

22 (22) *CAR: además /DCT era mucho libro /SCA para lo que +EMP [efamcv01d-56]

23 (23) *JOR: eh /TMT sumo diez //COM [efamcv08-48]

24 (24) *JOR: tú xxx //UNC [efamcv08-86]

3. Estructura y características del mini-corpus

La estructura este corpus está calcada de la estructura de los corpus del italiano y del portugués brasileño de IPIC (ver Tabla 6) para garantizar su comparabilidad. El español se ha organizado en 37 sesiones (cada sesión corresponde a una

grabación[3] transcrita y alienada con Winpitch) repartidas casi de modo homogéneo en tres tipos de interacción según el número de participantes (monólogo[4], diálogo y conversación) y según el contexto o situación en el que se desarrolla (Familiar/privado y Público). En la tabla 5 se ve también el número de secuencias prosódicas terminadas TS correspodientes.

Tabla 5.

Mini-corpus español				
Contexto de la interacción	Tipo de interacción	Sesiones	Palabras	TS
Familiar/ Privado	**mn** monólogos	7	7941	915
	dl diálogos	7	7200	1345
	cv conversaciones	8	8014	1693
Público	**mn** monólogos	3	3151	362
	dl diálogos	7	8179	1203
	cv conversaciones	5	6101	929
Totales		**37**	**40586**	**6447**

Algo más de la mitad de Mini-corpus procede de C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente)[5] (Nicolás Martínez 2012), la otra mitad son grabaciones transcritas buscando las situaciones y los temas que fueran semejantes a los que contenían los corpus italiano y portugués brasileño. Se ve en la tabla 5 la división en interacciones familiares, que son características de situaciones espontáneas, mientras las interacciones públicas se caracterizan por tratar temas previamente preparados y frente a personas no conocidas.

Otros requisitos importantes relacionados con la composición del corpus, y que también proceden del diseño de los corpus ilatiano y portugués brasileño, son:

- la máxima variación de temas representativos del habla espontánea (los temas de los tres corpus se han elegido para que sean semejantes por la situación y finalidad comunicativa;

[3] Un cierto número de sesiones corresponde a la misma grabación, en tales casos se añade una letra tras en nombre para identificar a la misma grabación en momentos sucesivos, ejemplo: efamcv01d.

[4] Se trata de situaciones en las que un solo hablante tiene prevalentemente la palabra.

[5] Corpus de libre acceso consultable en: http http://lablita.it/app/cordial/ descrito en Nicolás Martínez 2012.

- la presencia de hombres y de mujeres de distintas edades, que se ve reflejado en los metadados que preceden a cada sesión;
- habla sin variación diatópica, en el caso del corpus español se trata de hablantes de Madrid o que viven allí;
- la máxima espontaneidad del habla, para lograrlo la mayoría de las grabaciones se han hecho sin avisar previamente a los hablantes;
- el periodo en el que se han hecho las grabaciones ha sido entre los años 2004 y 2017[6];
- todas las grabaciones tienen alta calidad acústica.

Un aspecto importante de este corpus es su comparabilidad con los otros corpus italiano y portugués brasileño. Aunque no es el tema central de este artículo se puede observar en la siguiente tabla (Tabla 6) las características comunes de los tres corpus. A propósito de la cantidad de palabras equivalente en las distintas interacciones se puede observar que los monólogos públicos cuentan con menor número de palabras; como se ve esto sucede en los tres corpus.

Tabla 6.

CORPUS		**BRASILEÑO**			**ITALIANO**			**ESPAÑOL**		
	Int.	**S.**	**Palab.**	**TS**	**S.**	**Palab.**	**TS**	**S.**	**Palab.**	**TS**
Familiar / Privado	**mn**	6	8122	856	6	9463	1090	7	7941	915
	dl	5	8017	1877	5	7530	1786	7	7200	1345
	cv	4	6107	1408	3	4975	1286	8	8014	1693
Público	**mn**	1	1486	143	2	2730	265	3	3151	362
	dl	2	2916	584	2	3046	555	7	8179	1203
	cv	2	3261	645	2	4845	705	5	6101	926
Total		20	29909	5513	20	32589	5687	37	40586	6447

En la transcripción del Mini-corpus el etiquetado es para representar la segmentación del fluido del habla en unidades prosódicas. Se ha hecho

[6] Las grabaciones del los corpus de las otras dos lenguas no corresponden exactamente al mism operiodo.

identificando las lindes prosódicas perceptivamente relevantes basándose en el modelo de C-ORAL-ROM[7] (Cresti & Moneglia 2005) y C-Or-DiAL[8].
En el Anejo 1 se ve que las primeras seis etiquetas son las utilizadas para marcar las lindes prosódicas de las unidades prosódicas.

Las últimas ocho etiquetas descritas en el Anejo 1 se utilizan para mostrar la estructura del diálogo (<__>,#); las incidencias externas al discurso (hhh); sonidos o vocalizaciones que afectan al discurso (yyy, mh, eh, e'); fragmentos de palabras (&__) y fragmentos del audio que no se pueden transcribir o que se han querido no transcribir para mantener anónimo el discurso (xxx, HIDl).

4 Interfaz de BD-IPIC

En la interfaz[9] de DB-IPIC http://lablita.it/app/dbipic/ se ven numerosas áreas que ofrecen numerosas posibilidades combinatorias de extracción de información.

Las opciones de búsqueda que ofrece son (ver en el Anejo 2 Imagen 1):

- *Source selection:* permite la selección de uno de los corpus en una de las tres lenguas.
- *General filters:* ofrece dos tipos de elecciones: *Reference units filter* permite elegir uno de los tipos entre las unidades de referencia (*Utterance*, *Stanza*) y su tipo de composición (*simple* o *compound*) y *Metadata Filter* se elige la tipología de texto según el tipo de interacción: (*conversations*, *dialogues*, *monologues*) y según el contexto comunicativo (*familiar*/*private*, *public*).
- *Information Pattern:* permite hacer búsquedas en el área *Search for Information Pattern* que permite la elección de una unidad informativa que se quiera buscar en las partes del corpus arriba seleccionadas. Se puede también elegir la posición de la unidad que se busca (al principio o al final de la *Utterance*/*Stanza*).
- *Word restrictions*: sirve para búsquedas de palabras contenidas en la unidad elegida; en *Linear relation between selected units* se incluye o excluyen las unidades seleccionadas según su relación de contigüidad.

[7] Para consultar el etiquetado C-ORAL-ROM, que hoy se puede considerar establecido como tradicional para los corpus del italiano, del portugués y del español se puede ver: Moneglia & Cresti 1997.
[8] Ha habido pequeñas diferencias puesto que en C-Or-DiAL: se han ocultado nombres con un sonido marcado con xxx y no se ha usado HIDl; las disfluencias se han diferenciado en [/] repetición, [//] reformulación y [///] reinicio, sin usar [/n]; las vocalizaciones han sido marcadas con &mm y &eh.
[9] Trabajo descrito con detalle en Panunzi & Gregori 2011.

- *Utterance restrictions:* permite seleccionar restricciones que incluyan o excluyan etiquetas de la articulación, palabras y categorías gramaticales.

Los resultados de las búsquedas muestran (ver en el Anejo 2 Imagen 2) la transcripción con el etiquetado a los dos niveles, se puede abrir el audio correspondiente y en el caso de querer el contexto se puede ver la transcripción de toda la sesión (clicando sobre el nombre de la sesión) y obtener sus datos y metadatos; también es posible salvar en Excel los resultados de cada búsqueda.

4.1 Ejemplos prácticos: utilidad de las búsquedas

En este apartado se da una lista de búsquedas en IPIC para mostrar los múltiples tipos de estudio que se pueden servir para extraer información esta base de datos. Esta panorámica no es ni exhaustiva ni tan siquiera representativa del uso de IPIC, los comentarios que se han hecho son las primeras reflexiones que estos parciales resultados podrían invitar a hacer.
La finalidad de estos ejemplos de búsqueda es invitar a quien lea este artículo a utilizar esta herramienta, pues ofrece datos concretos del habla espontánea.

- Uso de *Metadata filters* y de *Word restrictions* para la comparación de algunos aspectos de las unidades Incipit INP y Conector Discursivo DCT:
  INP en monólogos 52 casos en interacciones 178. Se pueden estudiar los contextos con: 94 *bueno*, 23 *pues*, 21 *pero*, 12 *a ver*, 12 *y*, etc.
  DCT en monólogos 198 en interacciones 277. Se pueden estudiar los contextos con: 147 *y*, 98 *entonces*, 84 *o sea*, 60 *pero*, 30 *pues*, 22 *bueno*, etc.

Estos datos en una primera lectura invitarían a una comparación entre estas dos unidades, pues hay algunas palabras (*pero*, *pues*, *bueno*) que cumplen una función muy distinta al encontrarse en unidades distintas; se podrían estudiar cuáles son las otras palabras[10] que se presentan en estas unidades y también en qué proporción con relación a las otras palabras se usan.

- Uso de *Linear relation* para estudiar un contexto contiguo: Escansión SCA seguida de Paréntesis PAR se presenta en 8 contextos, mientras sin contigüidad se encuentra 58 veces.

[10] Sobre los marcadores es interesante ver el artículo de Pons Bordería 2018.

Esta búsqueda ofrece sin duda la posibilidad de iniciar una investigación sobre cuáles son las unidades que van antes de PAR y si el caso de SCA es especialmente significativo. Se tendría que ver si para SCA también es significativo que se anteponga a PAR o si hay otra unidad textual nuclear o no nuclear o dialógica a la que preceda de modo relevante.

- Uso de *Start* y *End of Utterance* para observar el uso del Fático PHA (342 concordancias totales) en: principio de enunciado 124, final 203, intermedio 15.

Se podría observar su función en las varias posiciones. En una primera observación se ve que tienen especial interés los siguientes enunciados:

(25) *PER: ya /[INP] bueno /[PHA] un parking te meterás aquí /[COB] dejarás el coche lo que < sea > //[COM] [efamdl01-139]

(26) *PIL: se ha declarado culpable /[SCA] de todos los cargos que le imputaban /[COB] y a cambio de eso < no le han > [/3] < no le van > [/3] vamos /[PHA] no le han condenado a [/1] a muerte //[COM] [efamcv06-122]

(27) *CAR: es mejor que sea un pan /[i-COM] ya /[PHA] mh sí /[PHA] duro / de +[EMP][efamdl02-92]

(28) *CAR: porque mira que me ha hecho fumar Juan /[CMM] eh /[PHA] y Chunita //[CMM] [efamcv04-50]

- Uso de la palabra *sí* en un Fático PHA en posición inicial 40, en final 14, en intermedia 3. Uso de *no* en PHA en posición inicial 37 (de estos 4 con *pero, pues, claro, sé*), final 127, medio 3. En total uso de *sí* y *no* en PHA 224 sobre 342 casos.

Este tipo de análisis comparativos y de la distribución de varias palabras en una determinada unidad ofrecen información muy interesante sobre sus características.

- Otras unidades en posición final son: El Apéndice de Comment APC: 177 final, 4 medio. A diferencia del Paréntesis PAR: 59 en posición final, en posción media 244.

Para la caracterización y redefinición en español de una unidad es muy interesante ver la posición que ocupan cuando esta información resulta significativa.

- El Tópic TOP puede llevar formas verbales en algunos casos: 296 sobre los 915 totales.

También puede ser interesante plantear estudios sobre los contenidos gramaticales de las unidades. Sobre el Tópic en particular los estudios son muchos el último de Cresti y Moneglia 2018.

Reflexiones finales

Este artículo presenta los apectos esenciales de la teoría L-AcT y las características principales del Mini-corpus del español de DB-IPIC.

Respecto a L-AcT se describen las unidades de referencia (enunciados y estancias) y su estructura informativa que se articula en un conjunto de funciones textuales y dialógicas.

Respecto al Mini-coprus del español se han mostrado sus características cuantitativas y cualitativas que han respetado las condiciones necesarias para la comparabilidad con los corpus italiano y portugués. Para ilustrar su utilidad de la base de datos DB-IPIC se han dado ejemplos de búsquedas.

Se ha subrayado cómo para asentar las bases de L-AcT se ha procedido de modo empírico, evitando interpretaciones de cualquier tipo, se ha tratado de un proceso de constatación de hechos lo que ha llevado a poner en evidencia los componentes de la estructura informativa. El hecho de que también en español se constaten las mismas unidades que en italiano y portugués refuerza su validez. Estudios comparativos sobre aspectos fundamentales de los procesos comunicativos podrían constatar variantes puntuales entre estas lenguas.

Esta base de datos y este Mini-corpus ofrecen, por lo tanto, un recurso para estudios sobre el español y sobre sus relaciones con otras lenguas.

Bibliografía

Austin, J.L. 1962. *How to do things with words*. Oxford: Oxford University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finnegan (eds). 1999. *The Longman Grammar of Spoken and Written English*, London: Longman.

Cresti, E. 2018. The illocution-prosody relation and the information pattern in the spontaneous speech according to the Language into Act Theory (L-AcT). En Moroni, M., Heinz, M.

(eds). *Linguistik Prosody: Grammar, information structure, interaction*. 88/1: https://bop.unibe.ch/index.php/linguistik-online/index

Cresti, E. 2000. Corpus *di italiano parlato*, 2 voll., CD-ROM. Firenze: Accademia della Crusca.

Cresti, E.. & M. Moneglia. 2018. The definition of the Topic within Language into Act Theory and its identification in spontaneous speech corpora. En Borreguero Zuloaga, M., V. Atayan & S. Grosse (eds), *Revue Romane*, vol. 53:1, 30-62.

Cresti, E. & M. Moneglia. 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam-Philadelphia: John Benjamins.

Cresti, E. & M. Moneglia. 2010. Informational patterning theory and the corpus-based description of spoken language. The compositionality issue in the topic-comment pattern. En M. Moneglia & A. Panunzi (eds), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Firenze: FUP.

Mittmann, M. M. & T. Raso. 2012. The C-ORAL-BRASIL Informationally Tagged Mini-Corpus. En Mello H., A. Panunzi & T. Raso (eds). *Illocution, modality, attitude, information patterning and speech annotation*. Firenze: Firenze University Press, 151-183.

Moneglia, M. & E. Cresti. 1997. Intonazione e criteri di trascrizione del parlato. En U. Bortolini & E. Pizzuto (eds), *Il progetto CHILDES Italia*, vol. II. Pisa: Il Cerro, 59-90.

Moneglia, M. & Cresti, E. 2006. "C-ORAL-ROM. Prosodic Boundaries for Spontaneous Speech Analysis." En *Spoken Language Corpus and Linguistics Informatics*, Kawaguchi, Y., S. Zaima & T. Takagaki (eds). Amsterdam: Benjamins, 89-112.

Nicolás Martínez, C. 2012. *C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente)*. Madrid: Liceus.

Panunzi, A., & L. Gregori. 2012. DB-IPIC: An XML database for the representation of information structure in spoken language. En Mello H. R., A. Panunzi & T. Raso (eds), *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*, Firenze: Firenze University Press, 133-150.

Panunzi, A. & M. Mittmann. 2014. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. En *Spoken Corpora and Linguistic Studies*, Raso, T. & H. Mello (eds), Amsterdam: Benjamins, 189-227.

Pons Borderia, S. 2018. The combination of discourse markers in spontaneous conversations. En Borreguero Zuloaga, M., V. Atayan & S. Grosse (eds), *Revue Romane*, vol. 53:1, 121-158.

Raso, T. & Mello, H. 2010. The C-ORAL-BRASIL corpus. En M. Moneglia & A. Panunzi (eds), *Bootstrapping Information from Corpora in a Cross-Linguistic Perspective*. Università degli Studi di Firenze, 193-213.

Raso, T. & Mello, H. (eds) 2012. *C-ORAL-BRASIL I. Corpus de referência para a fala espontânea informal do português do Brasil*. Belo Horizonte: UFMG.

WinPitch. http://www.winpitch.com/

Anejo 1

?	Delimita una secuencia prosódica autónoma con un claro perfil prosódico interrogativo.
…	Delimita una secuencia prosódica autónoma abandonada voluntariamente por el hablante con un perfil prosódico de suspensión.
//	Marca el final de todas las secuencias prosódicas independientes que no pertenecen a las tres clases anteriores. Indica la linde terminal (*terminal break*).
+	Señala una secuencia que ha sido interrumpida por el mismo hablante o por otro. El programa discursivo del hablante queda roto y la secuencia puede ser imposible de interpretar.
/	Señala una linde prosódica no terminal (*no terminal break*); delimita las unidades prosódicas.
[/n]	Indica la presencia de disfluencias (como falsos inicios o reformulaciones), n corresponde al número de palabras que el hablante ha desechado. Se puede considerar un tipo de linde prosódica no terminal, pero las palabras desechadas no contribuyen al modelo informativo o al contenido semántico de lo enunciado.
< >	Corchetes angulares que delimitan los solapamientos, es decir, los fragmentos de conversación en los que se superpone el habla de dos o más hablantes.
#	Etiqueta que indica un silencio considerable en medio de la conversación.
hhh	Etiqueta que indica la presencia de sonidos no pertenecientes al intercambio comunicativo, en general sonidos ambientales, como un timbrazo, una sirena de ambulancia, etc.
yyy	Etiqueta que indica la presencia de un sonido comunicativo que en un intercambio aporta algo a la comunicación, como una risa, un grito o un golpe para marcar énfasis, etc.
mh, eh	Etiqueta que indica la presencia de vocalizaciones más o menos breves.
&_	Etiqueta que precede a un fragmento de palabra, como: &telefo, &es.
xxx	Etiqueta que indica la ausencia de un fragmento de conversación imposible de entender y de transcribir.
HID	Etiqueta que sustituye a los nombres de persona o de entidades que se han cancelado para que permanezcan anónimas; la letra adjunta es para su identificación si son mencionadas en varias ocasiones.

Anejo 2

Imagen 1

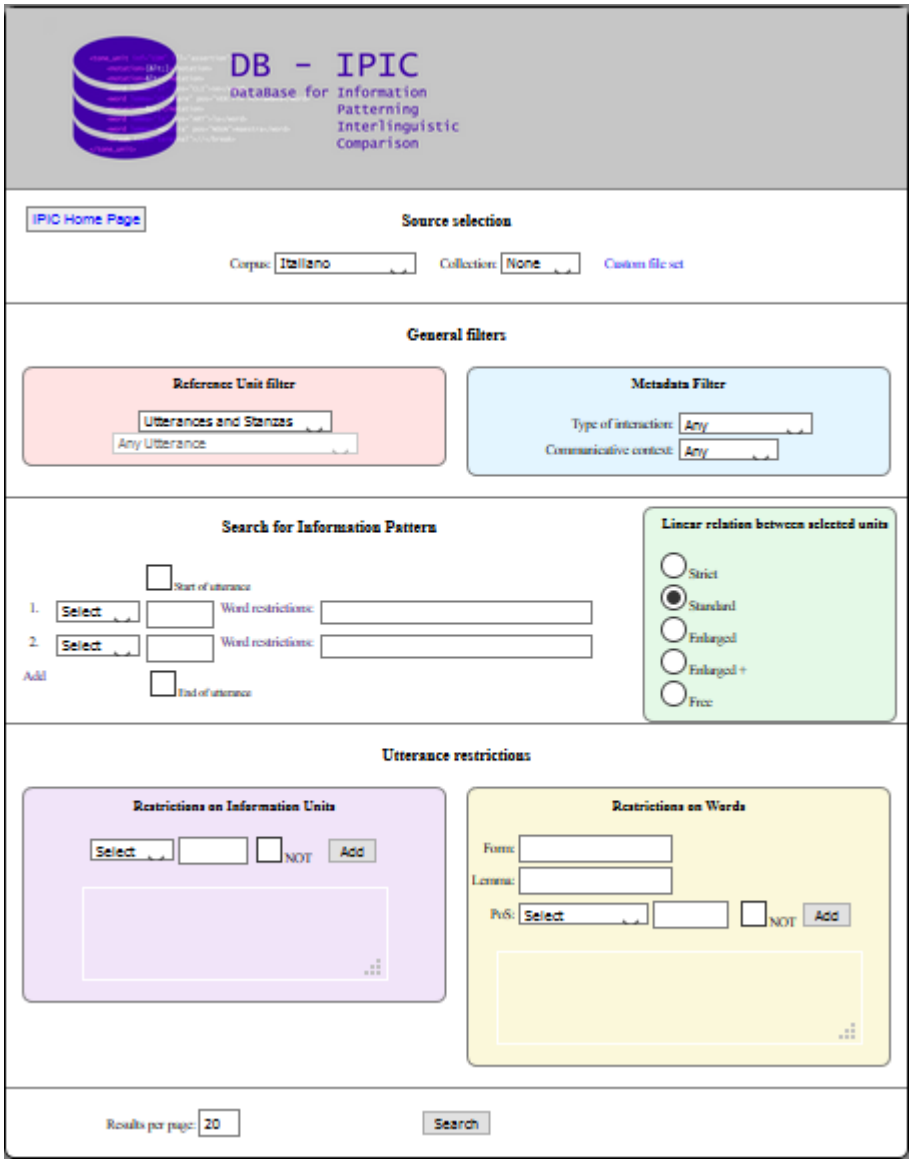

Imagen 2


XQuery
New Search
found 54 hits in 392 ms.
showing results 1 - 20
Next
sfamcv02a TOP COM ALL
CAR 29 entonces si no cabe todo / cómo hacemos / < hija > ?
sfamcv03 INP TOP COM
LOL 54 bueno / a ver si algún día / consigues sacarlo entero //
sfamcv03 TOP TOP COM
CAR 69 hhh pero era [/1] < era que a lo mejor / si la niña un día / no se > da mucha cuenta /
COM
y empieza de &rom [/1] yyy y se aficiona a romper < papeles > //
sfamcv04 INP TOP SCA COM
CAR 130 < pero > / si lo van a poner en dvd < ahora > / < lo vamos a parar > [/4] yo lo voy a parar / para hacer algunas fotos //
sfamcv07 OCT TOP COM
DIE 98 entonces / si hay gente comiendo / te largan //
sfamcv07 INP TOP PAR COM
MAA 195 a ver / si a lo tonto / ya en realidad / tengo bastante material //