

Developing a Pragmatic Annotation Scheme for a Spoken Corpus

María Sampedro Mella

Université catholique de Louvain, Belgium

maria.sampedro@uclouvain.be

Abstract

This paper presents a pragmatic annotation scheme for speech acts developed for the ESLORA corpus of spoken Spanish. The scheme integrates a pragmatic layer into an XML-based corpus infrastructure and encodes speech acts through an adjacency-pair-based model. In addition to enabling the retrieval of speech acts as independent categories, the system allows for queries that combine speech acts with lexical, morphosyntactic, and sociolinguistic variables. Given that corpus annotation has traditionally focused on lexical and morphosyntactic information (POS tagging), this proposed annotation scheme represents a valuable contribution to the field. The present paper describes the annotation scheme, XML implementation, query options, and limitations, thereby contributing to a computational model for corpus pragmatics.

Keywords: corpus pragmatics, pragmatic annotation, speech acts, spoken corpus.

1 Introduction

While linguistic corpora have been extensively used to analyse lexical and morphosyntactic phenomena, their use in pragmatic research remains more limited. This imbalance is due both to the influence of the grammatical tradition and to the part-of-speech (POS) tagging systems used to annotate corpora, which facilitate the retrieval of words, lexical combinations, categories, or lemmas. However, they usually do not encode pragmatic content such as speech acts, which makes it difficult to retrieve this type of data across registers or communicative situations.

Consequently, when using corpora in pragmatic research, researchers often have to perform queries based on representative forms (O’Keffe, 2018) or manually examine the materials of each corpus (Sampedro Mella, 2021a). This process is time-consuming and methodologically challenging, as

pragmatic phenomena are highly context-dependent, require the interpretation of speaker intentions (Weisser, 2015; Milà-García, 2018), and often rely on multimodal or paralinguistic cues (e.g., intonation, gesture, or prosodic emphasis). As a result, most corpus-based pragmatic studies rely on concordance searches of textual (i.e., transcribed) renderings of spoken discourse rather than on multimodal approaches (Adolphs & Chen, 2021). These limitations reflect broader methodological issues in the field of *corpus pragmatics*.

Even though corpus-based approaches have been used in pragmatic research for some time, corpus pragmatics is still a relatively novel discipline (Adolphs & Chen, 2021; Aijmer, 2026): “(...) corpus-pragmatic research is more than just pragmatic research and it is more than just corpus-linguistic analysis in that it integrates the horizontal (qualitative) methodology typical of pragmatics with the vertical (quantitative) methodology predominant in corpus linguistics” (Aijmer & Rühlemann, 2015: 12). Through computational procedures, this field seeks to contribute to the creation, enrichment, and pragmatic use of corpora.

Overall, pragmatic annotation has not been prioritised in corpus linguistics (Milà-García, 2018; Jucker, 2018), largely due to the linguistic and technical challenges that it entails. Annotating pragmatic phenomena is typically a manual process, since there is no one-to-one correspondence between speech acts and the linguistic or extralinguistic cues associated with them, and it often requires contextual interpretation: “Any type of linguistic annotation is a highly complex and interpretive process, but none more so than pragmatic annotation” (Weisser, 2015: 84). For this reason, in the context of spoken corpora, “pragmatic annotation is still much less advanced than other linguistic annotation types” (Rabadán, Ramón & Sanjurjo-González, 2021: 210).

Several English corpora containing pragmatic information have been developed over the past few decades. One of the first is the Switchboard Dialogue Act Corpus (Jurafsky, Shriberg & Biasca, 1997), which includes telephone conversations and is annotated with dialogue act tags. Similarly, the DialogBank (Bunt et al., 2016), developed under the ISO 24617-2 standard (Bunt et al., 2020), and the SPAADIA corpus (Leech & Weisser, 2013) provide structured dialogue act annotation frameworks for conversational data. The SPICE-Ireland Corpus (Kirk, 2013), by contrast, incorporates pragmatic annotation within a broader corpus-linguistic and prosodic framework.

In Spanish, only a limited number of corpora include pragmatic annotation systems that allow for the retrieval of speech acts, such as the C-ORAL-ROM corpus (Cresti & Moneglia, 2005), the C-Or-DiAL corpus (Nicolás Martínez, 2012), the CLANES corpus, and the COREMAH corpus (Vacas Matos, 2017). These corpora differ in both their design and their annotation schemes: the C-ORAL-ROM corpus focuses on spontaneous speech and includes prosodic annotation, although its pragmatic encoding is not primarily designed for the retrieval of speech acts; the C-Or-DiAL corpus is didactically oriented; the CLANES corpus is not publicly available, and the COREMAH corpus contains speech acts elicited through role-play interactions in which participants perform apologies, compliments, and negative responses to invitations.

The annotation schemes of these various corpora were developed for different interactional settings and research objectives. Furthermore, none of these corpora systematically allows for the extraction and categorisation of speech acts as recognised pragmatic units (see Sampedro Mella, 2026, for further information). For these reasons, these annotation schemes could not be directly incorporated into the annotation framework described in the present study.

In this context, the present paper introduces a pragmatic annotation scheme specifically developed for the ESLORA corpus. It encodes different speech acts in spoken interaction through a structured XML-based framework. By doing so, it aims to facilitate the retrieval of speech acts in spoken data without manual inspection. This contribution goes beyond XML annotation by providing a formal encoding of speech acts in

corpus data that enables their automatic extraction. Consequently, this annotation scheme facilitates systematic research in corpus pragmatics and can be adapted and applied to other spoken or written corpora and languages.

2 ESLORA Corpus

The ESLORA¹ (*Corpus para el estudio del español oral*) corpus is a spoken-language corpus designed to analyse Spanish in Galicia, a bilingual region in the north-west of Spain. Its current version (2.5) consists of audio recordings and transcriptions of 60 hours of sociolinguistic interviews and 24 hours of casual speech.

The data were manually transcribed and aligned using either Transcriber or ELAN tools (Vázquez Rozas & Barcala, 2020: 226). Following transcription, the texts were morphosyntactically annotated and lemmatised. The morphosyntactic annotation was implemented using an XML-based format to facilitate advanced linguistic searches and analyses. This annotation includes several features of spoken language, such as pauses, hesitations, overlaps, and repetitions (Vázquez Rozas & Barcala, 2020).

The pragmatic annotation scheme was implemented only in the interview subcorpus. Interviewees were selected according to key social variables including age, gender, and education level, to create a balanced corpus and facilitate variationist analyses.

Sociolinguistic interviews are often relatively formal, which may lead speakers to monitor their speech and reduce spontaneity — a limitation commonly described as the observer’s paradox (Labov, 1972). To mitigate this, ESLORA uses “free interviews” (Fernández Sanmartín, 2018, 2026) designed to approximate casual conversation. Rather than following a strict question-and-answer structure, interviewees can elaborate on their responses. Interviewers also seek to minimise the asymmetry of their role by contributing to the interaction with comments in a more dialogic way.

3 Methodology: pragmatic annotation and implementation

3.1 Scheme design

The annotation scheme was designed to enable the systematic identification and retrieval of speech

¹ <https://eslora.usc.es/search>

acts in the ESLORA corpus through its query tool. It was therefore necessary to adapt the scheme to the structure of the interactions and to corpus-specific constraints, such as transcript segmentation and the interactional characteristics of sociolinguistic interviews. As in other spoken corpora involving pragmatic annotation, the target categories were selected based on the type of data and the speech acts typically occurring in the corpus interactions. The selected categories include greetings, directive acts, offers, direct reported speech, and questions (see Appendix A.1), all of which can be identified and delimited with sufficient consistency for annotation purposes.

Given the nature of the sociolinguistic interviews, which, despite efforts to reduce formality and approximate casual conversation, still exhibit characteristics of traditional interviews, it was not possible to directly adapt and apply an existing annotation scheme developed for other corpora. These include, for instance, corpora of casual conversations (e.g., the CORDIAL corpus) or role-play interactions (e.g., COREMAH).

The speech acts targeted were organised into adjacency pairs, i.e., sequentially related pairs of turns produced by two different speakers, in which the first turn elicits the second (Schegloff & Sacks, 1973; Schiffrin, 1986). For example, a greeting or a question, as first-pair parts, elicits a return greeting or an answer, respectively.

Speech acts were manually annotated using an XML-based encoding system described in section 3.2. Three researchers were involved in the annotation and review process following an annotation guide². Any discrepancies were resolved through discussion, with consensus-based decisions adopted to ensure consistency in the application of the annotation scheme.

To assess the effectiveness of the encoding system, a series of evaluations of annotation consistency and information retrieval was carried out. In addition, the query system was tested by retrieving all annotated speech acts, confirming that they are consistently recognised as structured units.

3.2 XML encoding and annotation workflow

The speech-act taxonomy and annotation scheme were designed by the author and implemented in XML using a dedicated annotation

environment developed by NLPgo Technologies S.L. In the case of questions, due to their high frequency in a corpus of sociolinguistic interviews, a further subcategorisation was included. Specifically, the type of question had to be selected, such as yes/no questions, wh-questions, and echo questions, among others (see Appendix A.1).

The annotation scheme relied on explicit opening and closing XML tags, which were manually inserted at the beginning and the end of each speech act within the first-pair-part turn of the adjacency pair (e.g., a question or an offer) (see Appendix A.2). The same procedure was applied to the second-pair-part turn (e.g., the answer to the question or to the offer). In cases where no response was produced (e.g., unanswered questions), a “no response” placeholder was introduced to preserve the structural integrity of the annotation scheme³. Finally, when multiple speech acts co-occurred within a single turn (e.g., two questions or a combination of a directive and a question), each act was assigned a numerical identifier, which was consistently implemented across both first-pair-part and second-pair-part turns to avoid overlaps between annotations.

3.3 Query options

The ESLORA corpus query system allows for two main types of searches that can be further constrained using sociolinguistic filters, such as the speakers’ age, gender, education level, and role (interviewer, interviewee, or audience member).

The first search type (see Appendix B.1) retrieves speech acts and individual adjacency-pair components, i.e., first-pair parts or responses. Additionally, it allows for the simultaneous retrieval of multiple speech act types within a single query.

The second type of search (see Appendix B.2) enables more complex query configurations, in which speech acts can be combined with linguistic markers such as lexical or morphosyntactic information. This includes searches based on specific lemmas, word forms, or grammatical patterns. For instance, researchers can search for directive acts or offers combined with imperative verb forms, or direct reported speech combined with specific verb lemmas, allowing them to examine different persons and tenses. Such

² https://eslora.usc.es/guide_pragmatic_annotation

³ Note that direct reported speech does not conform to an adjacency-pair structure. It was annotated during the original

corpus transcription process and later incorporated into the set of speech acts retrievable through the query system.

functionality enables the systematic association between speech acts and their linguistic encoding.

Query results are displayed as concordance lists and can be exported in standard formats (e.g., CSV or Excel). Each result is also linked to its corresponding audio segment.

4 Scope and limitations

The main contribution of this pragmatic annotation scheme is the incorporation of a pragmatic layer into an already morphosyntactically annotated spoken corpus. Unlike traditional POS-based annotation systems, which primarily encode morphosyntactic or lexical information, this system enables direct access to speech act categories without requiring the manual inspection of transcripts. Furthermore, it is also possible to retrieve these acts alongside lexical, morphosyntactic, and sociolinguistic variables.

A second contribution lies in the adjacency-pair-based encoding strategy. This design not only identifies speech act categories but also retrieves their corresponding second-pair parts, i.e., responses within interactional sequences. In this respect, the scheme differs from other speech-act annotated corpora (e.g., C-ORAL-ROM, C-Or-DiAL, or CLANES). In those corpora, pragmatic information is not consistently structured to allow the retrieval of second-pair parts as complete interactional units.

The proposed pragmatic annotation system has significant applications in both research and in L2 teaching and learning. It facilitates the systematic study of speech acts in real interactional settings, enabling the analysis of their co-occurrence and variation according to sociolinguistic and discourse patterns. Moreover, the possibility of combining pragmatic information with morphosyntactic information allows for the study of the interface between grammar and pragmatics in spoken data. In L2 courses, the retrieved data can be used through data-driven learning (DDL) (Leech, 1997) to illustrate pragmatic functions in authentic communicative situations. In this way, it can complement classroom materials (see, e.g., Sampedro Mella, 2021b; Ballester de Celis & Sampedro Mella, 2021), thus contributing to the development of learners' pragmatic competence (Sampedro Mella & Rodríguez Domínguez, 2026).

As for limitations, the scope of the annotation scheme is currently limited to the sociolinguistic interviews of the ESLORA corpus. The speech act taxonomy focuses on a restricted set of categories

and does not incorporate prosodic information. Additionally, since the data consist exclusively of audio recordings and their transcriptions, the annotation does not include gestural or other types of multimodal information.

The annotation process also revealed several challenges. As is common in pragmatic annotation, the delimitation of speech acts was not always clear. In particular, difficulties arose in cases involving what Sacks, Schegloff, and Jefferson (1974: 711) describe as “expanded sequences”, where speech acts extended beyond adjacency pairs. This phenomenon was particularly frequent in the case of offers, as speakers often repeated or reformulated them, and interlocutors produced multiple refusals (see Appendix C.1). Similar difficulties were encountered in question-answer and directive act-response sequences. In these cases, a question or directive act was sometimes responded to with another question, which was subsequently answered, making it difficult to determine where the response to the initial speech act began (see Appendix C.2, and Sampedro Mella, 2026 for a more detailed discussion and examples).

Finally, although the annotation process involved a two-stage review procedure to ensure consistency, some degree of variability may remain. When pragmatic interpretation is highly context-dependent, especially when it comes to the identification or classification of some speech acts, annotators may arrive at different interpretations despite following the annotation guide. For instance, imperatives have generally been classified as directive speech acts (e.g., *cuéntame* ‘tell me’). However, they can also function as offers, as in invitations to come to the interviewees’ home or to take a seat (e.g., *siéntate* ‘sit down’). Such cases illustrate the inherent challenges of assigning categories to context-dependent pragmatic phenomena, highlighting the complexity of pragmatic annotation.

5 Conclusions and future work

This paper has presented the design and implementation of a pragmatic annotation scheme for speech acts in the ESLORA subcorpus of sociolinguistic interviews. The system was integrated into a corpus that already included POS annotation, enabling multi-layered analysis. The annotation scheme contributes to corpus pragmatics by making speech acts systematically retrievable as pragmatic units, both independently and in combination with (socio)linguistic variables.

In addition, the adjacency-pair-based encoding enables the retrieval of responses (second-pair-part turns), which broadens the possibilities of the corpus query system.

Future work should focus on expanding the taxonomy to other speech acts, refining the annotation guide for ambiguous cases, and incorporating prosodic information. Additionally, future research may explore adapting the scheme to other corpora and developing semi-automatic procedures for pragmatic annotation, including the potential application of AI-assisted approaches.

Building on this line of research, ongoing work is exploring the pragmatic annotation of other phenomena, such as address forms (i.e. *tú/usted* ‘you’), in the Corpus of Online Forum Replies (COFRE), a multilingual corpus of forum data from TripAdvisor. To this end, an automatic system based on spaCy has been developed to identify and classify address forms through pronoun- and verb-based detection. Future developments will focus on the validation and refinement of these automatic procedures, as well as on their integration into broader annotation schemes for multilingual corpus pragmatics (see Sampedro Mella et al., 2026 for further information).

Together, these developments contribute to addressing a broader challenge in corpus pragmatics: the limited availability of corpora enriched with pragmatic annotation. By making pragmatic phenomena systematically searchable and analysable, such resources open new possibilities for corpus-based pragmatic research and data-driven approaches to L2 teaching and learning.

6 Acknowledgements

The research presented in this paper represents an earlier stage of a broader research trajectory that is currently developed through the Marie Skłodowska-Curie Action (MSCA) *Personal and non-personal forms of address in three contiguous Romance languages: sociolinguistic approach and gender implications (DE-PERSONALISATION)*. (Grant code 101149488).

References

Svenja Adolphs and Yaoyao Chen. 2021. Corpus Pragmatics. In Michael Haugh Daniel Z. Kádár and Marina Terkourafi (eds.), *Sociopragmatics*, pages 639–662. Cambridge University Press.

Karin Aijmer. 2026. Corpus pragmatics. In Jesús Romero Trillo (ed.), *The Cambridge Handbook of Language in Context*, pages 289–313. Cambridge University Press.

Karin Aijmer and Cristoph Rühlemann. 2015. *Corpus Pragmatics: A Handbook*. Cambridge University Press.

Carmen Ballester de Celis and María Sampedro Mella. 2021. La didáctica del español como lengua extranjera: un acercamiento desde la lingüística de corpus. In Carmen Ballester de Celis and María Sampedro Mella (eds.), *Aportes del CAES a la enseñanza del español como lengua extranjera: estudios aplicados de gramática y pragmática*. Special issue of *Verba*, 81, pages 9–20.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang and Kars Wijnhoven. 2016. The DialogBank. In *Proceedings 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3151–3158.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. 2020. The ISO Standard for Dialogue Act Annotation, Second Edition. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 549–558.

CLANES Corpus - Controlled LAnguage English-Spanish. ACTRESS Group. <https://actres.unileon.es/wp/es/corpus-comparables/>

Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins.

ESLORA Corpus - *Corpus para el estudio del español oral*, v. 2.5. <https://eslora.usc.es/>

Alba Fernández Sanmartín. 2018. La entrevista libre como método para evitar la paradoja del observador. *CHIMERA*, 5(2): 141–196.

Alba Fernández Sanmartín. 2026. La entrevista sociolingüística en el corpus ESLORA: análisis pragmático-discursivo de la interacción entrevistador–entrevistado. *Variación. Revista de variación y cambio lingüístico*, 3: 204–228.

Andreas Jucker. 2018. Corpus pragmatics. In Andreas Jucker, Klaus Schneider and Wolfram Bublitz (eds.), *Methods in Pragmatics*, pages 455–466. De Gruyter.

- Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. *Switchboard SWBD-DAMSL Shallow Discourse-Function Annotation: Coders Manual*. Institute of Cognitive Science Technical Report. University of Colorado.
- John M. Kirk. 2013. Beyond the Structural Levels of Language: An Introduction to the SPICE-Ireland Corpus and its Uses'. In Janet Cruickshank and Robert McColl Millar (eds.), *After the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting*, pages 207–232.
- William Labov. 1972. Some Principles of Linguistic Methodology. *Language in Society*, 1(1): 97–120.
- Geoffrey Leech. 1997. Teaching and Language Corpora: A Convergence. In Anne Wichmann and Steven Fligelstone (eds.): *Teaching and Language Corpora*, pages 1–23. Routledge.
- Geoffrey Leech and Martin Weisser. 2013. *The SPAADIA Annotation Scheme*. https://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf
- Alba Milà-Garcia. 2018. Pragmatic annotation for a multi-layered analysis of speech acts: a methodological proposal. *Corpus Pragmatics*, 2(3): 265–287.
- Carlota Nicolás Martínez. 2012. C-Or-DiAL (*Corpus Oral Didáctico Anotado Lingüísticamente*). Liceus.
- Anne O’Keeffe 2018. Corpus-based function-to-form approaches. In Andreas H. Jucker, Klaus P. Schneider and Wolfram Bublitz (eds.), *Methods in Pragmatics (Handbooks of Pragmatics)*, pages 587–618. De Gruyter.
- Rosa Rabadán, Noelia Ramón and Hugo Sanjurjo-González. 2021. Pragmatic annotation of a domain-restricted English-Spanish comparable corpus. *Bergen Language and Linguistics Studies*, 11(1): 209–223.
- Harvey Sacks, Emmanuel A. Schegloff and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4): 696–735
- María Sampedro Mella. 2021a. DCT vs. corpus orales: reflexiones metodológicas sobre el estudio de los actos de habla. *Pragmalingüística*, 29: 337–395.
- María Sampedro Mella. 2021b. *Estimado Sr. vs. Hola hotel*. El análisis contrastivo de la interlengua para la enseñanza de la variación pragmático-discursiva. In Vicente Marcet Rodríguez, Vanesa Álvarez Rosa and Manuel Nevot (eds.), *La variación en español y su enseñanza. Reflexiones y propuestas didácticas*, pages 134–150. Universidad de Salamanca.
- María Sampedro Mella. 2026. Dando voz a la pragmática: creación de un sistema de anotación de actos de habla en el corpus ESLORA. *Variación. Revista de variación y cambio lingüístico*, 3: 39–67.
- María Sampedro Mella, Louise-Amélie Cougnon and Barbara De Cock. 2026. COFRE – Corpus of Online Forum Replies. A Trilingual TripAdvisor Corpus for the Study of Address Forms and Beyond. *Variación. Revista de variación y cambio lingüístico*, 3: 132–156.
- María Sampedro Mella and Ana Rodríguez Domínguez. 2026. Las funciones comunicativas en ELE: de los libros de texto a los corpus con anotación pragmática. In I. Fernández López (ed.), *Competencia plurilingüe y pluricultural: enfoques desde la Didáctica de la Lengua y la Literatura*, pages 385–403. Peter Lang.
- Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up Closings. *Semiotica*, 8(4): 289–327.
- Deborah Schiffrin. 1986. Functions of and in discourse. *Journal of Pragmatics*, 10(1): 41–66.
- Marta Vacas Matos. 2017. *COREMAH – Corpus Español Multimodal de Actos de Habla*. <https://coremah-1a37f.firebaseio.com/>
- Victoria Vázquez Rozas and Mario Barcala. 2020. Computational tools and spoken corpora design: an ongoing dialogue. *Caplletra*, 69: 221–240.
- Martin Weisser. 2018. *How to Do Corpus Pragmatics on Pragmatically Annotated Data*. John Benjamins.

Appendix A. Speech acts in the ESLORA corpus

A.1 Speech acts annotated in the ESLORA corpus

Speech act (first pair-part)	Description/examples	Responses (second pair-part)
Greetings / introductions	Greetings and introductions, occurring either separately or in combination.	Greetings or no response.
Directive acts	Requests, petitions, orders, and advice.	Responses to directive acts (acceptance, refusal, thanking) or no response.
Offer–Invitation	Proposals, offers, and invitations.	Acceptance, refusal, thanking, etc., or no response.
Direct reported speech	Quotations.	Ø
Questions		
Yes/no questions.	E.g., <i>¿viviste siempre aquí desde pequeño?</i> ‘Have you lived here since you were a child?’	Response or no response.
Wh-questions.	E.g., <i>¿a qué colegio fuiste?</i> ‘Which school did you go to?’	Response or no response.
Alternative questions.	E.g., <i>¿tienes buen recuerdo de los estudios o no?</i> ‘Do you have a good memory of the studies or not?’	Response or no response.
Echo questions.	Questions that repeat part of a previous utterance: <i>—es complicado ¿eh! — ¿complicado?</i> ‘—It’s complicated, huh! — Complicated?’	Response or no response.
Incomplete questions.	Question left unfinished that prompt the interlocutor to respond: <i>¿hay algo más que quieras contarme de ti de?</i> ‘Is there anything more you want to tell me about yourself, about...?’	Response or no response.
Tag questions.	Statements followed by a short question tag: <i>pero eso es muy peligroso ¿no?</i> ‘But that’s very dangerous, isn’t it?’	Response or no response.

A2. Example of speech act annotation (yes-no question) in XML. The speech act tags are displayed on the right.

The screenshot shows an XML editor interface. The main window displays the XML content for a speech act, with fields for 'Comienzo' (Start) and 'Fin' (End) times, and 'Hablante' (Speaker). The XML code includes tags for the start and end of the speech act, as well as the speaker. The right sidebar shows a list of speech act tags, including 'a_acto_directivo_a_inicio', 'a_acto_directivo_b_fin', 'a_acto_directivo_respuesta_a_inicio', 'a_acto_directivo_respuesta_b_fin', 'a_ofrecimiento_a_inicio', 'a_ofrecimiento_b_fin', 'a_ofrecimiento_respuesta_a_inicio', 'a_ofrecimiento_respuesta_b_fin', 'a_pregunta_a_inicio', 'a_pregunta_b_fin', 'a_pregunta_respuesta_a_inicio', 'a_pregunta_respuesta_b_fin', 'a_saludo_a_inicio', 'a_saludo_b_fin', 'a_saludo_respuesta_a_inicio', and 'a_saludo_respuesta_b_fin'.

Appendix B. Corpus query tool for speech acts⁴

B1. Search for speech acts. Steps 1 and 2:

The first screenshot shows the search configuration interface. The 'Búsqueda' section has 'Corpus' set to 'Cualquiera' and 'Tipo' set to 'Palab. ortográficas'. The 'Sensibilidad' section has 'Acentos' and 'Mayúsculas' checked. The 'Resultado' section has 'Tipo de resultado' set to 'Intervenciones'. The 'Ordenación' section has 'Coincidencia' selected. The 'Tamaño de página' is set to 50. The 'Desde' and 'Hasta' fields are set to 'Cualquiera'. The 'Papel' field is set to 'Cualquiera'. The 'Nivel de estudios' field is set to 'Cualquiera'. The 'Sexo' field is set to 'Cualquiera'. The 'Habla' field is set to 'Cualquiera'. The 'Buscar en' field is set to 'Cualquiera'.

The second screenshot shows the search results for 'Intervenciones'. The 'Tipo de intervención' is set to 'Cualquiera'. The results list includes: 'Saludo / Presentación', 'Saludo / Presentación inicio', 'Saludo / Presentación respuesta', 'Acto directivo', 'Acto directivo inicio', 'Acto directivo respuesta', 'Ofrecimiento', and 'Ofrecimiento inicio'.

B2. Search for speech acts in combination with linguistics units

The screenshot shows the search configuration interface. The 'Búsqueda' section has 'Corpus' set to 'Cualquiera' and 'Tipo' set to 'Elem. gramaticales'. The 'Sensibilidad' section has 'Acentos' and 'Mayúsculas' checked. The 'Resultado' section has 'Tipo de resultado' set to 'Concordancias'. The 'Ordenación' section has 'Coincidencia' selected. The 'Tamaño de página' is set to 50. The 'Desde' and 'Hasta' fields are set to 'Cualquiera'. The 'Papel' field is set to 'Cualquiera'. The 'Nivel de estudios' field is set to 'Cualquiera'. The 'Sexo' field is set to 'Cualquiera'. The 'Habla' field is set to 'Cualquiera'. The 'Buscar en' field is set to 'Cualquiera'. The 'Elem. gramaticales' section has 'Elemento gramatical' set to 'Etiqueta'. The 'Palabra ortográfica' field is set to 'Lema'. The 'Buscar' button is highlighted.

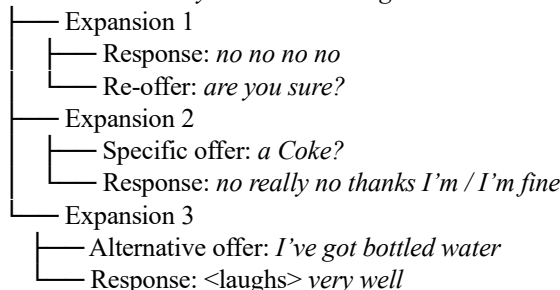
⁴ Source: Sampedro Mella (2026).

Appendix C. Examples of speech acts

C1. Offer with expanded sequences

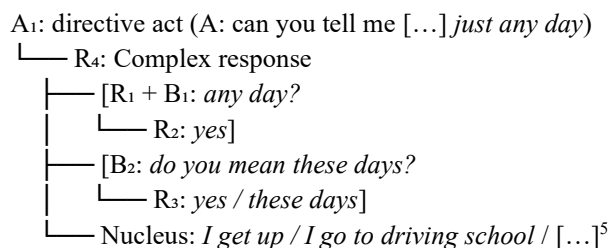
Interviewee (B): *eeh (1) ¿quieres algo de tomar? /*
‘uh (1) would you like something to drink? /’
Interviewer (A): *no no no no /*
B: *¿de verdad que no? (2) ¿una Coca-Cola? //*
‘are you sure? (2) a Coke?’ //
A: *no de verdad de verdad no gracias gracias estoy / estoy bien ¡eh! /*
‘no really really no thank you thank you I’m / I’m fine, really!’ /
B: *(3) tengo agua de botella bueno /*
‘I’ve got bottled water anyway /’
A: *<risa/> muy bien /*
‘<laughs/> very well /’ (SCOM_H31_046)

Main offer: *would you like something to drink?*



C2. Directive acts/questions with expanded sequences

Interviewer (A): *bueno/ me puedes contar por ejemplo qué haces un día normal en tu vida// un día cualquiera//* [directive act]
‘well/ can you tell me for example what you do on a normal day in your life// just any day’
Interviewee (B): *¿cualquiera?* [response to the directive act + yes–no question]
‘any day?’
A: *sí* [response to the yes–no question]
‘yes’
B: *¿de ahora mismo?* [yes–no question]
‘do you mean these days?’
A: *sí / de ahora mismo //* [response to the yes–no question]
‘yes / these days’
B: *me levanto / voy a la autoescuela / después voy por ahí con los amigos [...]*
‘I get up / I go to driving school / then I go out with my friends [...] (SCOM_H11_047)



⁵ Both diagrams were created using ChatGPT.