Presenting SULEC and SUNCODAC: Two Corpora for English Language Teaching and Research

SPERTUS (Spoken English Research Team at the University of Santiago de Compostela)

Ignacio Palacios (Coordinator), Rosa Alonso, Mario Cal Varela, Susana Doval Suárez, Ana Fernández Dobao, Javier Fernández Polo, Lidia Gómez García, Elsa González Álvarez, Paula López Rúa, Luisa Roca Varela, and Raquel P. Romasanta

II Xeira CLARIAH-GAL

SULEC The Santiago University Learner of English Corpus

What is SULEC?

The ***Santiago University Learner of English Corpus*** (SULEC) contains written and spoken data produced by students of English at three different proficiency levels: beginner, intermediate, and advanced. It comprises samples from 1,374 students, totaling 406,690 grammatical elements and 365,030 words.

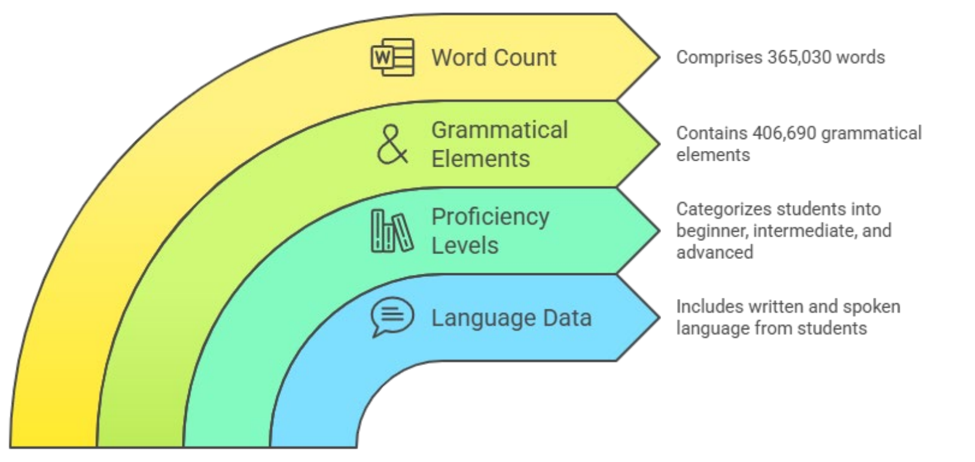

Word Count
Comprises 365,030 words
Grammatical Elements
Contains 406,690 grammatical elements
Proficiency Levels
Categorizes students into beginner, intermediate, and advanced
Language Data
Includes written and spoken language from students

Figure 1. SULEC corpus overview

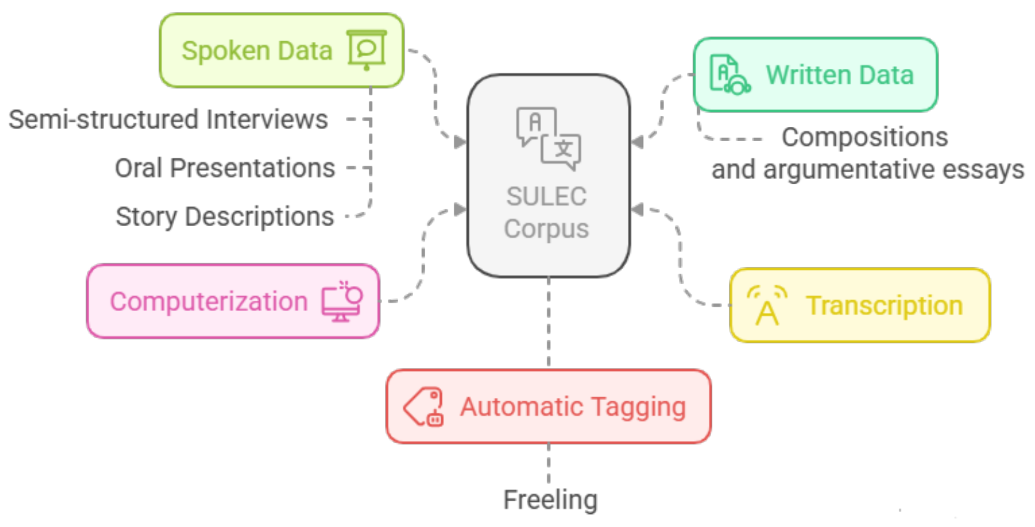


Figure 2. SULEC compilation process

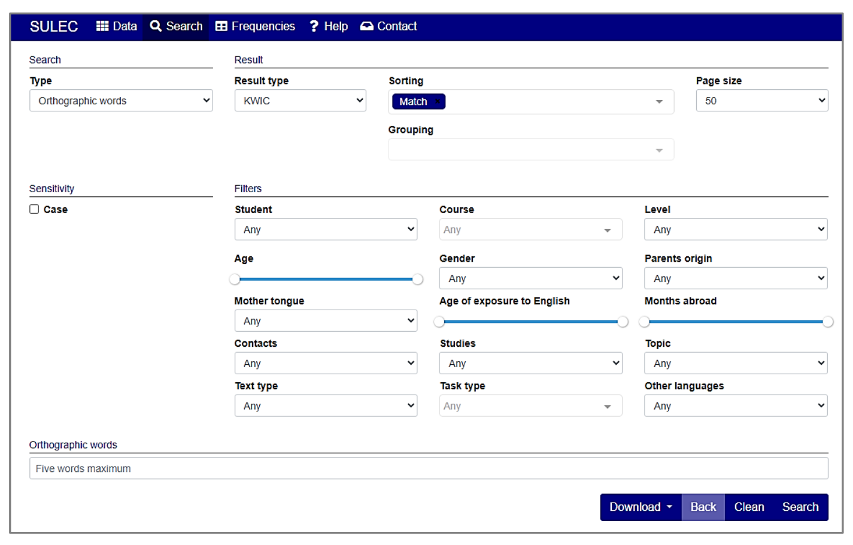
For the written component, students from Secondary Education and University (undergraduates in English Studies, Education, and Translation) submitted compositions and argumentative essays on topics such as the obligatoriness of military service, the value of the monarchy in Spain, and their views on recent anti-smoking policies, among others.

For the spoken component, data were drawn from students' presentations and oral tasks in which pairs of learners described a series of comic strips to construct a story.

All the data were transcribed and computerized to be fully automatically tagged using Freeling.

The interface (Figure 3) is user-friendly, allowing users to search for orthographic words and grammatical elements, through detailed fields, including tags and lemmas.

The **Search** section lets users choose a search type via the "Type" selector, including options for orthographic words, grammatical elements, and proximity-based searches. The **Result** section has four settings: *Result type* (with options like samples, frequencies, KWIC, and matching expressions), *Sorting*, *Grouping*, and *Page size*. **Sensitivity** allows users to set case sensitivity. **Filters** let users refine results using variables like mother tongue, proficiency level, gender, and age.



Search
Type: Orthographic words

Result
Result type: KWIC
Sorting: Match
Grouping
Page size: 50

Sensitivity
□ Case

Filters
Student: Any
Course: Any
Level: Any
Age
Gender: Any
Parents origin: Any
Mother tongue: Any
Age of exposure to English
Months abroad
Contacts: Any
Studies: Any
Topic: Any
Text type: Any
Task type: Any
Other languages: Any

Orthographic words
Five words maximum

Figure 3. The SULEC interface

What are the applications of SULEC?

1. Teaching:

Teachers of English can use numerous examples of learner language in the classroom. They can highlight common errors or showcase effective writing and oral performance from the corpus.

2. Second Language research:

- **Phonological:** Common pronunciation difficulties.
- **Morpho-syntactic:** Issues connected with word order, agreement, acquisition of specific structures (e.g. negatives, clefts).
- **Lexical:** Collocations, "false Friends", and concordances.
- **Discourse:** Organization of information, cohesive devices, discourse markers, and communication strategies.

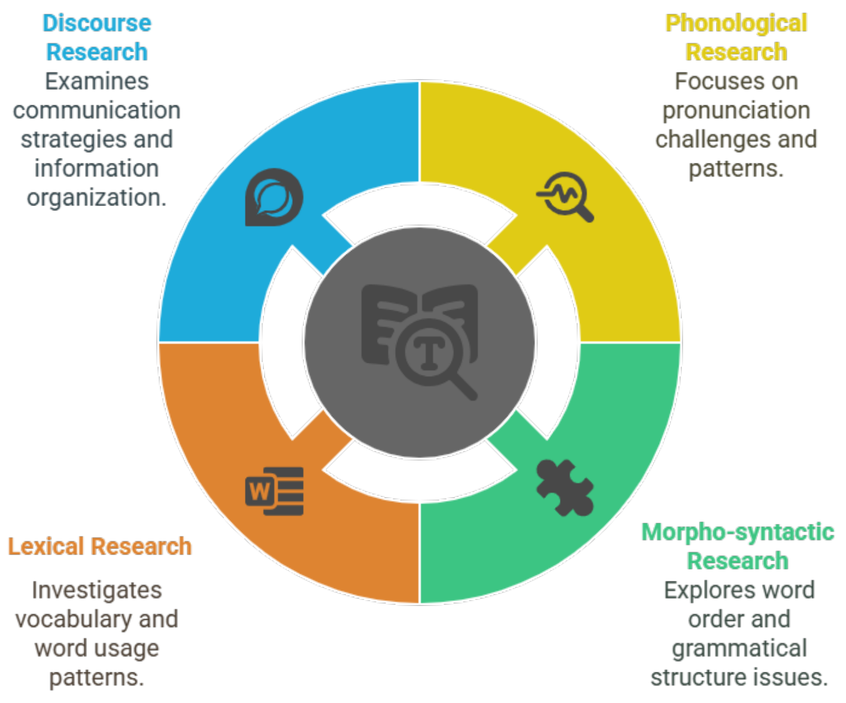

Discourse Research
Examines communication strategies and information organization.
Phonological Research
Focuses on pronunciation challenges and patterns.
Lexical Research
Investigates vocabulary and word usage patterns.
Morpho-syntactic Research
Explores word order and grammatical structure issues.

Figure 4. Conducting Second Language research on SULEC

3. Materials development:

Materials designers can draw on the corpus for examples or insights, for instance, identifying grammatical structures that pose challenges, or issues related to language transfer and interference.

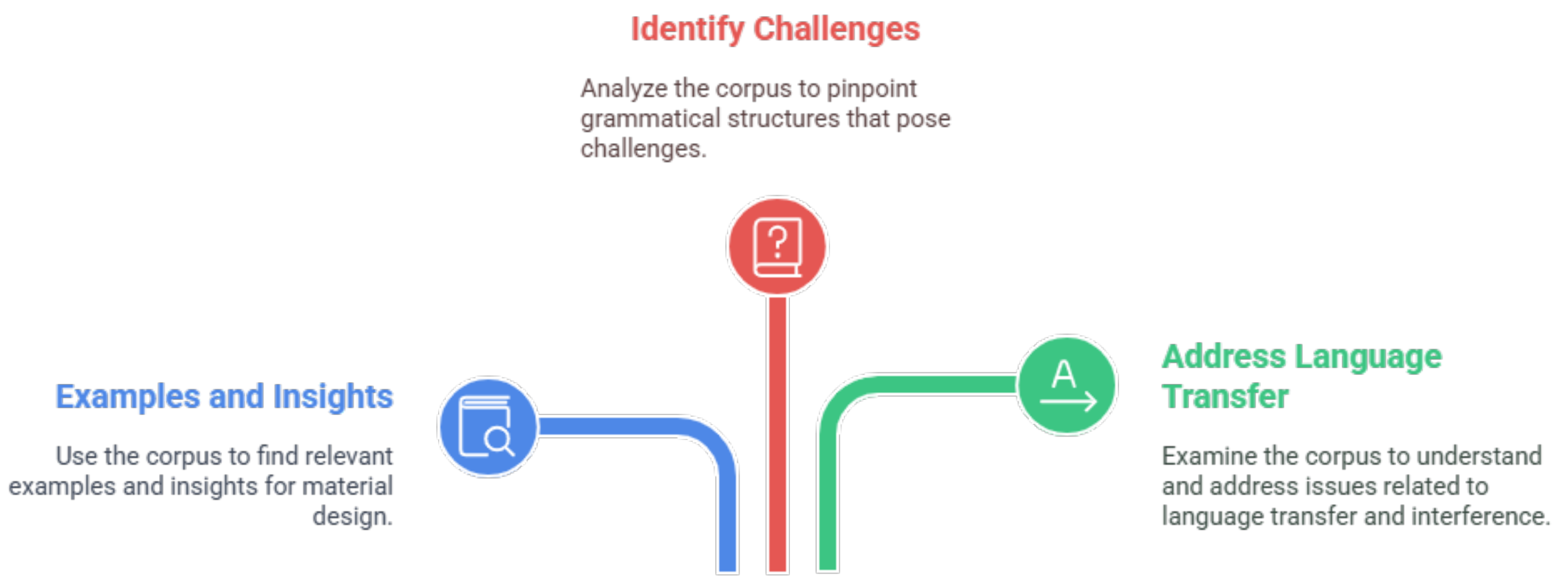

Identify Challenges
Analyze the corpus to pinpoint grammatical structures that pose challenges.
Examples and Insights
Use the corpus to find relevant examples and insights for material design.
Address Language Transfer
Examine the corpus to understand and address issues related to language transfer and interference.

Figure 5. Using SULEC for materials development

Some examples of published research on SULEC

- Fernández Dobao, Ana M. 2005. *The Use of Communication Strategies by Spanish Learners of English. A Study of the Collaborative Creation of Meaning, Language and Linguistic Knowledge*. PhD Thesis. Universidade de Santiago de Compostela.
- Roca Varela, M. Luisa. 2012. *New Insights into the Study of English False Friends: Their Use and Understanding by Spanish Learners of English*. PhD Thesis. Universidade de Santiago de Compostela.
- Roca Varela, M. Luisa. 2015. *False Friends in Learner Corpora. A Corpus-Based Study of English False Friends in the Written and Spoken Production of Spanish Learners*. Linguistic Insights Series. Bern: Peter Lang. ISBN 978-3-0343-1620-0.

SUNCODAC Santiago UNiversity Corpus Of Discussions in Academic Contexts

What is SUNCODAC?

The ***Santiago UNiversity Corpus Of Discussions in Academic Contexts*** (SUNCODAC) is a compilation of online discussions via Moodle. The participants were undergraduate students enrolled in a translation course which combined in-class interaction with collaborative work through the virtual learning environment. The largest portion of the messages contains students' comments and feedback on the initial translation proposal posted by a classmate. Posts containing the proponents' final summary and the lecturers' feedback are also included. Figure 6 shows the usual sequence of message types in each discussion thread.

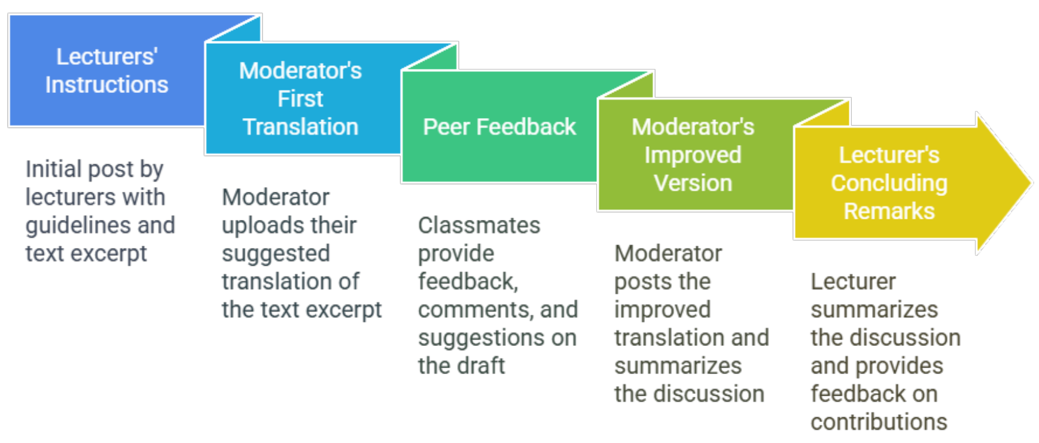

Lecturers' Instructions
Initial post by lecturers with guidelines and text excerpt
Moderator's First Translation
Moderator uploads their suggested translation of the text excerpt
Peer Feedback
Classmates provide feedback, comments, and suggestions on the draft
Moderator's Improved Version
Moderator posts the improved translation and summarizes the discussion
Lecturer's Concluding Remarks
Lecturer summarizes the discussion and provides feedback on contributions

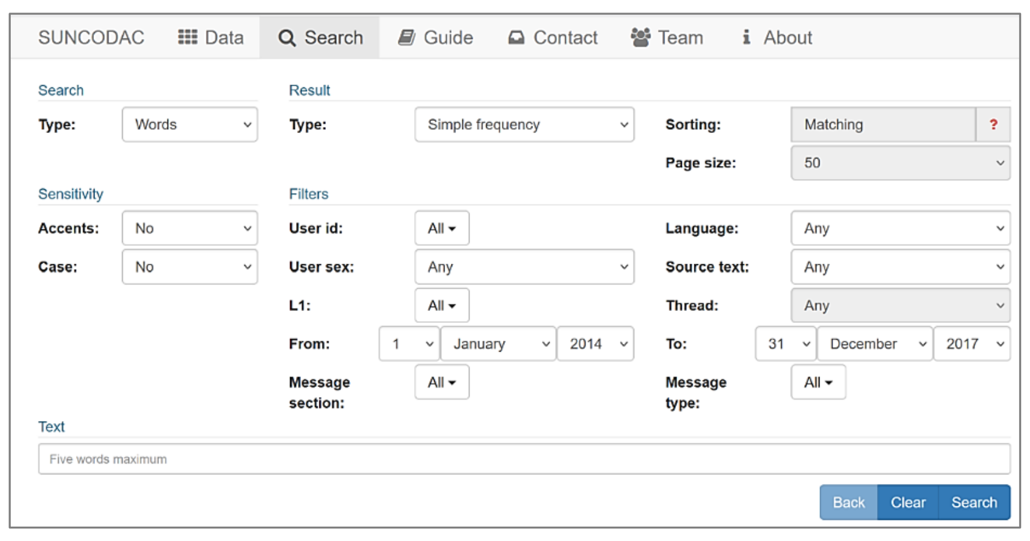
Figure 6. Schematic view of message types in each discussion thread

Language	Number of posts	Number of words
English	1,665	322,834
Spanish	1,521	232,440
Galician	119	18,547
Total	3,305	573,821

Table 1. Corpus holdings

The texts were collected over four consecutive academic years (2014-2017). Table 1 summarizes the corpus holdings. 520 students from various national backgrounds contributed to the corpus: local students (73.8%), Chinese students (14.6%), native English students (4.8%) and students from other language and national backgrounds, mostly European (6.73%).

Authorship and personal references were anonymized by replacing participants' names with a unique code. All posts were stored in xml format for exploration by means of an ad hoc corpus tool (see Figure 7), which allows users to retrieve full messages and reconstruct discussion threads, or to search for words and word sequences to explore language patterns.



Search
Type: Words

Result
Type: Simple frequency
Sorting: Matching ?
Page size: 50

Sensitivity
Accents: No
Case: No

Filters
User id: All
Language: Any
User sex: Any
Source text: Any
L1: All
Thread: Any
From: 1 January 2014
To: 31 December 2017
Message section: All
Message type: All

Text
Five words maximum

Figure 7. The SUNCODAC interface

What are the applications of SUNCODAC?

1. Research:

- **Sociolinguistics**: norm development, gendered language, style mixing, native vs non-native discourse, English as a Lingua Franca, classroom discourse;
- **Computer-mediated communication**: emoticons and paralinguistic cues;
- **Genre analysis**: move structure and characteristic language of student discussions;
- **Pragmatics**: hedging, indirectness, humour and other politeness strategies; cultural differences in pragmatic norms;
- **Discourse**: constructive feedback, expressing disagreement, constructing identity, manifesting solidarity, engaging in persuasion, metadiscourse;
- **Contrastive studies**: advice, politeness, criticism, etc. in L1 and L2 contexts; culture-specific practices.

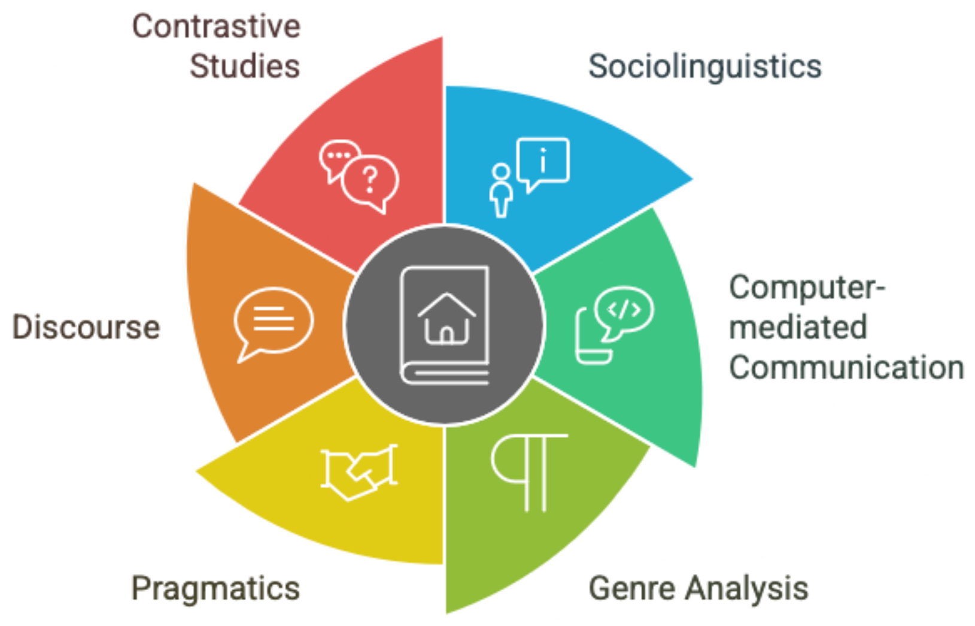


Figure 8. Conducting research on SUNCODAC

2. Teaching:

SUNCODAC may be used to teach and learn participation skills in collaborative learning contexts, particularly in English as a second language. It provides authentic materials for reflection on effective and flawed production. Grounding learning in real-world practice, instead of abstract rules, helps students develop practical skills and understand the principles of effective communication.

Possible uses of SUNCODAC might include:

- **Teaching material development**: based on observed errors, targeted to specific needs and learners;
- **Data-driven learning activities:** students explore effective and problematic language patterns.

SUNCODAC may be used by teachers and learners to deal with issues of:

- **Rhetoric**: persuasion, argument construction, identity creation.
- **Discourse**: post structure, metadiscourse, interpersonal skills.
- **Pragmatics**: speech acts: suggestions, criticism, etc.; politeness: hedges, disclaimers; humour and other rapport strategies;
- **Vocabulary and phraseology** realizing specific discourse functions.

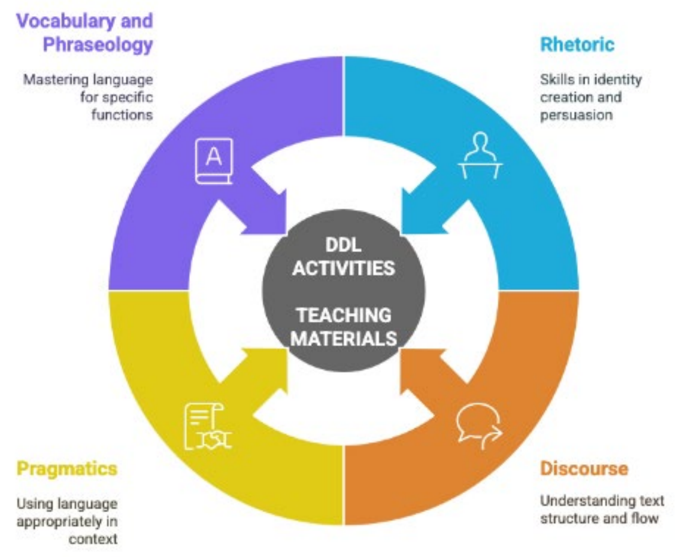

Vocabulary and Phraseology
Mastering language for specific functions
Rhetoric
Skills in identity creation and persuasion
DDL ACTIVITIES
TEACHING MATERIALS
Pragmatics
Using language appropriately in context
Discourse
Understanding text structure and flow

Figure 9. Enhancing language learning with SUNCODAC

Some examples of published research on SUNCODAC

- Cal Varela, M., & Fernández Polo, F. J. (2020). Preparing the ground for critical feedback in online discussions: A look at mitigation strategies. In J. Longhi & C. Marinica (Eds.), *CMC Corpora through the Prism of Digital Humanities* (pp. 15–34). L'Harmattan.
- Cal-Varela, M., & Fernández-Polo, F. J. (2022). Referring to other participants in asynchronous online discussions: Citation patterns in a higher education context. *Psychology of Language and Communication*, *26*(1), 353–374.
- Doval-Suárez, S. & González-Álvarez, E. (2025). "You have done a great job, but I would make some changes". Concession and politeness in asynchronous online discussion forums. *Research in Corpus Linguistics*, *13*(1), 113–138.

A complete description of the corpora, together with some tips to explore its contents can be found at the tools' websites:

SULEC

SUNCODAC

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN
USC SPERTUS
SPOKEN ENGLISH RESEARCH TEAM AT THE UNIVERSITY OF SANTIAGO DE COMPOSTELA
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
ILTIUS Instituto de Investigacións Lingüísticas e Literarias da USC
XUNTA DE GALICIA
CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL