SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: IS1312
STSM title: Testing the interoperability of annotation systems for oral DRDs in Spanish language
STSM start and end date: 15/01/2018 to 03/02/2018
Grantee name: Elena Pascual Aliaga

PURPOSE OF THE STSM:

The aim of the STSM carried out at the Université Catholique de Louvain under the supervision of Prof. Liesbeth Degand was to extend the annotation of spoken DRDs to the Spanish language, combining the proposals set out by Crible and Degand (2017a) and the Val.Es.Co. group (Briz and Pons 2010, Briz and Val.Es.Co. group 2014). The specific objectives of the stay were (1) to implement a merged protocol that includes both systems (see previous work by Crible and Pascual 2017, Pascual and Crible 2017); (2) to revise and expand the number of annotated DRDs in Spanish; (3) to discuss the problematic issues and questions that arise from applying a common annotation proposal; and (4) to deliberate on general issues related to the annotation of oral DRDs in spontaneous conversations.

During the stay we decided to focus on the application of Crible and Degand's annotation system to the Spanish data. Given that Crible and Degand's annotation system is still under development, we decided it would be more productive to concentrate exclusively on the application of this system with the aim of improving it. Objectives (1) and (3) were left aside as further steps to be developed at a later date in preparation for the final TextLink action conference in Toulouse, where the results obtained from the combination of both models will be presented.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

The stay commenced with the training of the grantee through background reading on the application of the "independent domain and function annotation scheme for spoken language" by Crible and Degand (2017a) and on general methodological aspects of corpus annotation. In this first phase of the three-week stay meetings were held in order to provide guidance on the application of Crible and Degand's annotation system.

The second phase of the stay was dedicated to the annotation of DRDs in a sample of three spontaneous Spanish conversations (12000 words approximately) from the *Corpus Val.Es.Co. 2.0* (Cabedo and Pons 2013). A total of 1399 DRD tokens were annotated. Concurrent meetings were held to discuss the results of the annotation, which included suggestions for the implementation of various aspects of the annotation protocol.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

1399 tokens of Spanish DRD were analysed and annotated once by a single annotator. A database containing all the annotations was created in Excel with the aim of carrying out a quantitative analysis of the annotation results. The results obtained from the annotation facilitated a series of theoretical reflections on the identification of DRD in Spanish and on the application of Crible and Degand's functional taxonomy.

In relation with the identification of DRD in Spanish, the definition followed was the one provided by Crible (2017), although some exceptions were made to favour a more inclusive approach. For example, response signals ("claro", "vale"), interjections ("¿eh?", "¡ah!", "tío"), some adverbs ("luego", "en realidad", "además", "a lo mejor") and general extenders ("y eso", "y tal") were included in the DM category and were annotated. Other elements such as *dicendi* verbs ("digo", "dice") or epistemic parentheticals ("se ve que", "la verdad es que", "seguramente", "parece que") were excluded.

The features that were annotated in the three conversations, following the proposal by Crible (2017) and the most recent revision by Crible and Degand (2017b), were the lexical item identified as a DRD, its domain and its function. The annotation of the domain and function was carried out independently, according to the most recent revision by Crible and Degand (2017b):

- DOMAINS: ideational (IDE), rhetorical (RHE), sequential (SEQ), interpersonal (INT).
- FUNCTIONS: addition (ADD), alternative (ALT), cause (CAU), concession (CONC), condition (COND), consequence (CONS), contrast (CONT), punctuation (PUNCT), temporal (TEMP) and specification (SPE).

In one of the three conversations, other features were tagged in addition to the previous ones, following the annotation scheme by Crible (2017):

- POS: the Part of Speech to which the DM corresponds, which may be CC (coordinating conjunction), RB (adverb), VP (verbal phrase), SC (subordinating conjunction), WP (pronoun), JJ (adjective), NN (noun phrase), PP (prepositional phrase), UH (interjection).
- CO-OCC: the co-occurrence of DM – defined in terms of the syntagmatic adjacency of two or more DM – was also tagged.
- TYPE: the type of DM in reference the relationship it establishes among textual units, which may be REL (for relational DM), NREL (for non-relational DM) or B (for both relational and non-relational DM).

The application of Crible and Degand's functional taxonomy was found to be useful for tagging all the tokens of DRD in the data. The distribution of the annotated domains and functions is shown in the following figures:

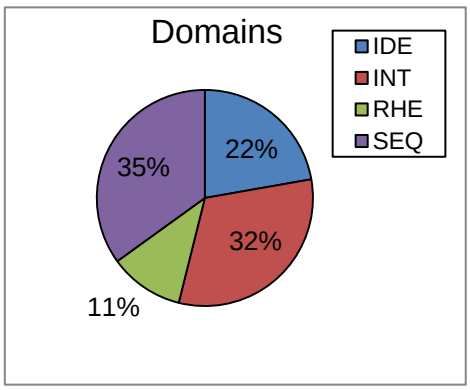

Figure 1: Distribution of domains

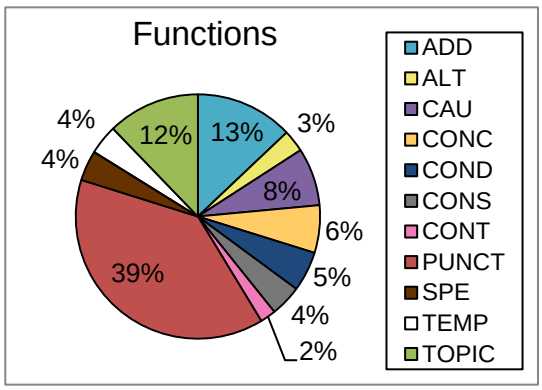

Figure 2: Distribution of functions

With respect to the relationship among domains and functions, the following table shows the most frequent functions found in each domain. The table has been refined and some values that are thought to be mistakes made during the annotation procedure have been removed:

IDE	RHE	SEQ	INT
ALT, CAU, CONC, COND, CONS, CONT, SPE, TEMP	ALT, CAU, CONC, COND, CONS, SPE	ADD, PUNCT, SPE, TOPIC	CONC, PUNCT,TOPIC

Table 1: Relationship among domains and functions

The use of Crible and Degand's functional taxonomy led to a number of reflections on the challenges faced in the application of the taxonomy to Spanish data:

- Some functions, even if considered independently with respect to the domains, are inevitably related to certain domains.
- The function PUNCT (39% of the DRD) may be subdivided into more specific functions in the interpersonal domain (for example, it might be useful to subdivide this function into one or two of the categories used in Crible 2017 – agreement/disagreement, face-saving, etc.).
- A number of issues related to the disentanglement of several functions and domains, such as:
  - Contrast and concession;
  - Punctuation and topic (especially in markers that seem to function in the interpersonal and sequential domain at the same time);
  - Alternative and specification.

FUTURE COLLABORATIONS (if applicable)

Possible avenues for future collaboration include the application of a common proposal bringing together Crible and Degand's system with the Val.Es.Co. system. Work on a common annotation proposal is already underway in work by Crible and Pascual (2017) and Pascual and Crible (2017), and future collaboration will capitalize on the preliminary findings of this research. The results obtained from the application of this merged annotation protocol will be presented at the final TextLink action conference in Toulouse.

The STMS has contributed to strengthening the links between two of the research groups and their respective annotation proposals for oral DRDs within the TextLink Action framework. The stay has set the foundations for further contrastive studies and collaborative work that will promote mutually enhancing dialogue between these two annotation proposals and, at the same time, facilitate the cross-linguistic analysis of DRDs across different annotated corpora.

REFERENCES

Briz, A. and Pons, S. (2010): "Unidades, marcadores discursivos y posicion", in O. Loureda and E. Acin (eds.): *Los Estudios sobre Marcadores del Discurso*. Madrid, Acro Libros, 523–557.

Briz, A. and Val.Es.Co. group (2014): "Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial)". *Estudios de Lingüística del Español*, 35(1), 11–71.

Cabedo, A. and Pons, S. (2013): *Corpus Val.Es.Co. 2.0* [online: www.valesco.es].

Crible, L. (2017): *Discourse Markers and (Dis)fluency across Registers. A Contrastive Usage-Based Study in English and French*. PhD Thesis. Université Catholique de Louvain.

Crible, L. and Degand, L. (2017a): "Independent domain and function annotation scheme for spoken language". *Corpus Linguistics and Lingustic Theory*. Advanced access: https://doi.org/10.1515/cllt-2016-0046.

Crible, L. and Degand, L. (2017b): "Testing interdependent annotation levels for sense disambiguation in

spoken English, French and Polish". In: 50th Annual Meeting of the Societas Linguistica Europaea, Zurich.
Crible, L. and Pascual, E. (2017): How to be (dis)fluent in English, French and Spanish: discourse markers within repetitions and repairs across languages. In: 15th International Pragmatics Conference, Belfast.
Pascual, E. and Crible, L. (2017): "Discourse markers within (dis)fluent constructions in English, French and Spanish casual conversations: The challenges of contrastive fluency research". In: Fluency and Disfluency Across Languages and Language Varieties, Louvain-la-Neuve.