Annotation of lexical bundles with discourse functions in a Spanish academic corpus

Eleonora Guzzi
Margarita Alonso-Ramos
Marcos García
Marcos García-Salido
Outline

1. Introduction
2. Goals
3. Methodology
   - Dataset
   - Annotation procedure
4. Results
5. Discussion
6. Conclusions
Introduction

**LEXICAL BUNDLES**

Academic English (Hyland, 2008, Simpson-Vlach and Ellis, 2010, etc.).

Recurrent lexical sequences with **high frequency** and **dispersion**, their linguistic value comes from the **discourse function** that they fulfill.

*it should be noted* (‘to emphasize’)
*as can be seen* (‘to resend’)
*it is clear that* (‘to show certainty’)

**MASTERY OF LBS**

The mastery of these LBs is crucial in academic writing.

**LEXICAL RESOURCES**

A few lexical resources in Spanish are available to offer aid to novice writers.
In our approach (MTT; Mel’čuk, 2015) **MWe** (or **phrasemes**) include compositional and non-compositional phrases that are associated to a discourse function.

They work as a whole and cannot be replaced by synonymous expressions that are unnatural:

- ✔️ f.i. (English) to put it differently
- ☹️ to use some different expressions
- ☹️ to say it in a different way

**Introduction: object of study**
Goals

Annotate 996 formulae assigned to 39 different discourse functions in a Spanish academic corpus.

Obtain a Spanish gold-standard corpus (1,800,000 words) for:

- Training and evaluating computational models to identify automatically LBs in new corpora.
- Linguistic analysis about the role of LBs in academic discourse.
Methodology

Dataset: corpus

HARTA-Exp; García-Salido et al. (2019)
- 2,025,092 word tokens
- 413 Spanish research articles (SERAC -Pérez-Llantada, 2008-)
- 4 main areas: (i) Arts and Humanities, (ii) Biology and Health Science, (iii) Physical Science and Engineering, (iv) Social Sciences and Education

- Tokenisation and lemmatization with LinguaKit
- PoS-tagged with FreeLing
- Dependency parsing (universal dependencies) with UDPipe
Dataset: formulae

985 formulae: recurrent sequences of words that are relevant for Spanish academic writing and fulfil a discourse function

E.g.

- to help writers to reformulate what is said: 'dicho de otro modo' (‘in other words’)
- to indicate opposition: 'no obstante' (‘however’)
- to express certainty: 'es sabido que' (‘it is well known that’)

Dataset: formulae

Identification using a semi-automatic method (García-Salido et al., 2018):

- **Automatic extraction** of 5,772 LBs (2-6 n-grams) from HARTA-Exp
- **Frequency** (≥10pmw) and **distribution** (≥1 in each area) threshold
- **Revision** of LBs by lexicographers to identify relevant academic formulae
- Assignation to the a **discourse function** based on García-Salido et al. (2019) classification.
Dataset: formulae

Classification

39 discourse functions → 3 groups

1. Bundles related to the research process: e.g. 'podemos concluir que' 'we can conclude that'
2. Text-oriented bundles: e.g. 'en primer lugar' 'first'
3. Interpersonal bundles: e.g. 'tal vez' 'perhaps'

Ambiguous formulae: they were assigned to the most frequent function
Annotation procedure

- **Pattern-matching in HARTA-Exp**
- **HARTA-Exp splitting and distribution among annotators**
- **Corpus validation by annotators**

- 985 formulae with discourse function tags
- Individual annotations + joint annotations

Inception

Gold-standard corpus
15 blocks of texts

HARTA-Exp splitting and distribution among annotators

5 + 1 for ANNOTATOR 1
5 + 1 for ANNOTATOR 2
5 + 1 for ANNOTATOR 3

CURATION: Consensus annotation
Se trata de una estrategia apuntada ya en otros ámbitos (Krauskopf y Vera 1995; Rojas 1998; Giménez et al. 2001a; Ramos 2001) y que va dirigida a aumentar la competencia entre los trabajos enviados por los autores potenciales del campo, con el beneficio añadido de atraer a más autores interesados en publicar sus originales en las revistas españolas debido al aumento de la calidad experimentado con dicha medida.

Comparación de la calidad formal

La situación global de la calidad formal de las publicaciones periódicas del campo de CCAFID en los años 2000 y 2005, así como la evolución sufrida entre dicho periodo, puede observarse en la tabla 3 en la que se muestra las puntuaciones de cada revista en el GGN y GFN, los indicadores de segundo orden elaborados a tal efecto.

Tabla 3: Comparación del GGN y del GFN de las revistas en 2000 y en 2005

La comparación de los valores medios del GGN de las revistas pone de manifiesto una mejora global de las publicaciones periódicas del campo durante los cinco años que median entre un análisis y otro.
Results & Discussion

- Annotated corpus of ca. 1,800,000 words (88% of HARTA-Exp)
- 360,000 words of consensus annotations

**INTER-ANNOTATOR AGREEMENT**

*Raw agreement:* 89% - 92% (positive overview)

*Krippendorff α* (Krippendorf, 2011): $\alpha=0.885$ - $\alpha=0.925$

**MANUAL EXAMINATION**

- 180 hours (individual annotations)
- Average of 414 changes per ca. 3,858 tagged formulae in each block

Linguists’ contribution has been essential
Types of changes:

1. Formulae that in some specific contexts they were not associated to any discourse function

Occurrences of 12 formulae were discarded

"Es más, la misma alumna emplea este apelativo dirigiéndose a un amigo o amiga."
'What is more, the student uses this appellation for addressing to a friend.'

"[...] debido a que su fabricación es más sencilla."
'[...] because its fabrication is more simple.'
Ambiguous formulae that are associated to two discourse functions

27 ambiguous formulae
E.g. substitution of ‘introduce the topic’ for ‘delimiting’ (499 times)

"De acuerdo con Takaday & Lourenço en 2004, las características generales de esta disciplina [...]"

'According to Takada % Lourenço in 2004, general features of this discipline [...]'.

"[...] tiene que ver con estrategias de actuación de cada biblioteca de acuerdo con su particular circunstancia local."

'[...] it has to do with strategies of action of each library according to their particular local circumstance.'
2 formulae that are nested but only the longest one was automatically tagged. Annotators selected the most relevant one.

como podemos observar en la tabla  ('as we can see in the table')

como podemos observar en (4-gram)  en la tabla (3-gram)

en ('in') belongs to both formulae
Occurrences of new formulae as different morpho-syntactic forms of existing ones

Morpho-syntactic variants of already registered ones.
11 different types of morpho-syntactic variants were added to the initial list of 985 --> 996 formulae.

*por una parte, por otra parte* (‘on the one hand, on the other hand’)

*abbreviated and grammatically correct counterpart: por una, por otra* (lit. ‘on the other’)*
Conclusions

- **Automatic techniques** used to identify vocabulary from corpus and for annotating formulae in corpora are a good starting point
- However, identification and annotation procedures still needed a **human validation**

**Ambiguity** is present:
- some LBs are formula in some contexts but not in others
- formulae that are associated to 2 discourse functions

Further work aims to use the gold-standard corpus obtained from this study to train and evaluate computational models that are capable of identifying automatically adequate LBs in new corpora, and for lexicographic and linguistic studies.
Thank you!

Acknowledgments

This research was funded by MCIN/AEI (PID2019-109683GB-C21 and PID2021-128811OA-I00), by Xunta de Galicia (ED431C 2020/11 and ED431F 2021/01), and by FEDER GALICIA 2014-2020 (ED431G 2019/01 and ED431G 2019/04). EG is funded by the Programa de Axudas á Etapa predoutoral da Xunta de Galicia, and MG by a Ramón y Cajal grant (RYC2019-028473-I).