Compilation and Annotation of Adjective-Adverb Interfaces in Romance

Towards a multilingual Open Access Corpus

Katharina Gerhalter, Martin Hummel, Gerlinde Schneider, Christopher Pollin
University of Graz

The project *Open Access Database: Adjective-Adverb Interfaces in Romance* aims at the creation of an annotated and lemmatised corpus of various linguistic phenomena related to Romance adjectives with adverbial functions. In this paper, we will explain the currently ongoing process of data compilation as well as the morphosyntactic and semantic categories for a thorough annotation by means of some Spanish examples.

Keywords: adjective-adverbs, Open Access, annotated corpus, pan-Romance corpus, historical corpus

1. Introduction

Over the last two decades, the research group on Adjective-Adverb Interfaces in Romance, located at University of Graz, has conducted several research projects entailing approximately 60 publications[1]. The research group focuses on various linguistic phenomena related to adjectives with adverbial functions: adjective-adverbs, such as Spanish *volar alto* / French *voler haut* 'to fly high' or Spanish *ver claro* / French *voir clair* 'to see clear'; adjectives used as discourse markers, such as Spanish *cierto* 'true'; and adverbial prepositional phrases including adjectives, for example *de seguro* 'certainly', *en serio* 'seriously, *a malas* 'badly, in bad terms', etc.

The long-term perspective has brought to light some problems concerning open access, sustainable storage and efficient usage of the analysed linguistic data.

[1] https://adjective-adverb.uni-graz.at/en/research/publications/

CHIMERA. Romance Corpora and Linguistic Studies 5.2 (2018), 305-311. ISSN 2386-2629

Labour-intensive updating of the databases can be sustainably guaranteed only if (i) not persons but institutions ensure the access and if (ii) international standards are created and implemented. As the research group cooperates with several international partners, who use and add data, the data should be tagged in a way that idiosyncratic solutions are reduced to a minimum.

Therefore, the objective of the project *Open Access Database - Adjective-Adverb Interfaces in Romance*[2] is the creation of a corpus for several Romance languages, where adjective-adverbs are uniformly and comprehensibly annotated and lemmatised. The project aims at documenting historical as well as present-day language examples. It updates already analysed and partially tagged subcorpora and further includes newly tagged data by the project team and by cooperation partners. The project is funded by the pilot program *Open Research Data* of the *Austrian Science Fund* (FWF: ORD 66-VO). Martin Hummel, the project leader, and Katharina Gerhalter, both from the Department of Romance Studies, take on the data collection and the elaboration of linguistic categories for the annotation. Additionally, Gerlinde Schneider and Christopher Pollin, both from the *Centre for Information Modelling – Austrian Centre for Digital Humanities*[3], located at University of Graz, are in charge of the data modelling, e.g. the annotation tool and the processing and displaying of the data. The duration of the project is set for two years (2017-2019). The final objective is to explore reasonable ways to make the collected and annotated linguistic research data openly accessible and reusable via a search mask.

2. Data compilation of Romance Adjective-Adverbs

Research on Romance adverbs traditionally focuses on those ending in *-ment(e)*. In contrast, less attention has been payed to adjective-adverbs, and even less to prepositional phrases. Adjective-adverbs are the only pan-Romanic deadjectival adverbs; their oral tradition leads directly back to Latin. Adjective-adverbs have largely been marginalized by normative standardization pressure towards *mente*-adverbs; therefore, they tend to appear more productively in substandard and regional varieties (Hummel 2017: 19-23).

The general underrepresentation of adjective-adverbs in historical corpora which are restricted to written sources, as well as the formal and functional interfaces between the traditional word classes *adjective* and *adverb* (Hummel & Valera 2017), challenge the compilation of examples. Unlike *mente*-adverbs, which

[2] https://adjective-adverb.uni-graz.at/en/research/projects/open-access-database-2017-2019/
[3] https://informationsmodellierung.uni-graz.at/en/

are unambiguously marked by the suffix *-mente*, no specific sequence can be digitally searched to obtain adjectives with adverbial functions.

For French, the Corpora *Frantext*[4] and *Corpus of the Dictionnaire du Moyen Français*[5] have been explored to compile the database of the *Dictionnaire historique de l'adjectif-adverbe* (Hummel & Gazdik in preparation). It contains over 13,000 examples from the 11th to the 20th century, which correspond to combinations of 700 different verbs with 300 different adjective-adverbs (e.g. *voir grand* 'to think big'). For the *Open Access Database*, all examples were annotated and lemmatized concerning the verb phrase *verb + adjective-adverb*. This historical data has been completed by a database labelled *documentation complémentaire* which includes approx. 4,500 examples of adjective-adverbs in present-day informal Internet usages, for example in blogs or discussion forums.

Lemmatized corpora such as the Spanish *Corpus del Diccionario Histórico* (=*CDH*) offer specific search combinations. Despite providing a categorization of word classes, the *CDH* does not classify adjective-adverbs systematically as "adverbs". Therefore, it is, for example, not possible to simply search for *justo* as an adverb. It is necessary to read examples of the adjective *justo* or to search for certain combinations that favour adverbial usage, before manually classifying adverbial uses (Gerhalter 2018: 47). By combining the most frequently used adjective-adverbs (lemmas) with the word class *verb* in the *CDH nuclear*, we collected and tagged approx. 2,200 examples of modal adverbs, such as *ver claro*. This database covers the 13th until the 21st century.

The old-fashioned but still effective and thorough approach of reading whole texts has been applied by Hummel (2014) for the *Sintáxis Histórica III*-chapter "Adjetivos Adverbiales". This *SH3*-database consists of approx. 1,200 examples from the 13th to 21st century. Not being restricted to the combination of *verb + adjective-adverb*, i.e. manner adverbs, it covers a wider range of syntactic functions (including Discourse Markers) as well as formal variation, such as inflected adverbs and prepositional phrases.

Currently, the Spanish database is being extended to include a systematic compilation of adjective-adverbs found via lemma search in the *Corpus Diacronico y Diatópico del Español de América* (=*CORDIAM*). This corpus includes examples from the 16th to the 19th century, especially from colonial administrative and juridical documents as well as chronicles and private letters. After searching for adjectival lemmas such as *seguro* we select and register all records of adverbial uses as well as the corresponding prepositional phrases (e.g. *de seguro*).

[4] http://www.frantext.fr/. Data retrieval from 2003-2005.

[5] http://www.atilf.fr/dmf/; version DMF1, 2003. Data retrieval from 2003-2005.

Adverbial phrases containing the pattern *preposition + adjective* (*de seguro, por cierto, al justo…*) are the focus of the current pan-Romanic project *The Third Way*[6] directed by Martin Hummel at the Romance department of the University of Graz (for the theoretical background, see Hummel in print). Examples are collected, tagged and analysed for several Romance languages and will also be integrated into the *Open Access Database of Adjective-Adverb Interfaces.*

2.1. Morphosyntactic and semantic categories for annotation

In order to lemmatize and classify the several functions and meanings of adjective-adverbs, we use an annotation tool. The lemmatization unifies orthographic variation–especially regarding historical data–and enables the search via lemmas. It further allows analysis of type-token-frequencies.

Additionally, every example is tagged with several categories. In the first place, the morphosyntactic classification takes into account the formal structure of the adverbial (e.g. adjective-adverb, *mente*-adverb or prepositional phrase), as well as its possible inflection. Furthermore, we assign several pre-defined categories for both the syntactic scope of the adverbial (*verb, verb and subject, verb and object, adjective, adverb, noun or phrase without verb-reference, sentence*) and its semantic classification (*manner, quantity/intensity, time, location, specification, discourse*). To illustrate these categories, we will cite five examples.

Example (1) shows a record of *ver claro* 'to see clearly'. The adjective-adverb *claro* is a manner adverb whose scope is the verb form *veo*:

(1) Cuando usted habla de la política del Ejército, hay algo que no veo *claro*. (1967, Viñas, David; Los hombres de a caballo; CDH)

In example (2), the scope of *altos* aims both at the verb *subir* and at the subject *los fumos*, and its semantic classification is those of *location*. The adjective-adverb shows plural-concordance (inflection) with the subject of the sentence:

(2) este pujamiento dell agua que fuera tanto en alto porque tan *altos* subieran los fumos de los sacrificios que los de Caím fizieran a los ídolos (1252-1284; Alfonso X; General Estoria. Primera Parte; p. 55, SH3)

In contrast, *justo* in (3) is a focus adverb. Its scope aims at the nominal phrase *un mes* and its semantic category is *specification*:

[6] https://adjective-adverb.uni-graz.at/en/research/projects/the-third-way-2018-2021/

(3) Hoy hace *justo* un mes, ¡oh suerte dura, / qué cerca está del bien la desventura! (1578, Ercilla, Alonso de; La Araucana, 2ª parte; CDH)

In (4), the adjective-adverb *harto* modifies the adjective *gustoso* (its scope) and semantically indicates intensification:

(4) te doy El Rey Gallo, y discursos de la Hormiga, plato *harto* gustoso y moral; creo que no te cansará (1671; Santos, Francisco; El rey gallo y discursos de la hormiga; p. 86, SH3)

Finally, in example (5), the adverbial phrase *de seguro* has a discourse function and its scope is the whole sentence:

(5) Es una propiedad valiosa que *de seguro* ha de tener muchos interesados. (1879; Documentos informativos; Uruguay; CORDIAM)

The tool further enables us to register modification of the adverb (e.g. *tan* in *tan altos*, example 2), coordination of two different adverbials (e.g. *hablar claro y alto*), as well as reduplication of the same type (e.g. *claro*, *claro…*).

In addition, the tool offers categories to tag and lemmatize verbs and to tag the subject of the sentence. Insofar, the search mask also will consider syntax (that is, word order). In the case of prepositional phrases, such as *a mis solas* or *al vivo*, prepositions, articles and possessives can be tagged.

3. Data Modelling and Processing

Over the course of several research projects, we have developed an integrated annotation model that combines all morphosyntactic and semantic annotations. The common reference model represented by a domain-specific RDFs ontology is also used for data retrieval and processing (Pollin *et al*. 2018).

The tagging and lemmatizing of the data are carried out manually by experts in (historical) Romance linguistics using an annotation tool implemented as an add-in for Microsoft Word. The reason for this is to provide a low-threshold data acquisition scenario. The add-in generates data encoded in XML/TEI[7] which is validated against a schema that implements the annotation model. All data storage, processing and analysis is based on this TEI-encoded data.

[7] www.tei-c.org

The data are archived and published through the certified repository GAMS (*Asset Management System for the Humanities*)[8]. To offer an appropriate long-term preservation and provision for the research data, the repository infrastructure is currently extended to provide genuine support for linguistic corpus data.

In the spirit of Open Access, the annotated data will be accessible in different data formats via interfaces such as those defined by the *European Research Infrastructure Consortium for Language Resources*[9]. These are the TEI/XML data itself, a highly structured RDF dataset as well as the TCF[10] format. In order to guarantee discoverability and reusability, a detailed description of the metadata for each sub-corpus will be provided via a CMDI[11] interface.

4. Outlook and forthcoming pan-Romance data

The *Adjective-Adverb Interfaces* Corpus will be divided into multiple sub-corpora of the individual Romance languages, as they show parallelisms: for example, Spanish and Portuguese *ver claro*, French *voir clair*, Italian *vedere chiaro* and Romanian *a vedea chiar*/*clar* (Chircu 2014). Therefore, in addition to the already mentioned French and Spanish data and the pan-Romanic project on prepositional phrases, the data compilation and annotation will also include cooperation with international project partners for Portuguese, (Old) Romanian and Italian (especially southern dialects) data.

Based on the annotation model, a search mask will be offered for in-depth search queries. It will allow general requests for lemmas, complex requests for the annotated morphosyntactic and semantic categories as well as combinations of various search criteria. Therefore, the pilot search mask from a previous project[12] is going to be adapted for the new infrastructure. For the purpose of scientific reuse, the results are offered as XML, Excel and PDF export files.

To sum up, a cross-linguistically applicable model for the annotation of adjective-adverbs is to be developed. This will also allow for the integration of new databases in the future. To the extent that this project is based on the idea of Open Science, we also encourage researchers in the field of Romance adverbs to annotate and integrate their data in the *Adjective-Adverb Interfaces* corpus.

---

[8] http://gams.uni-graz.at

[9] https://www.clarin.eu

[10] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

[11] https://www.clarin.eu/content/component-metadata

[12] https://adjective-adverb.uni-graz.at/en/databases/adjective-adverb-database/

References

Academia Mexicana de la Lengua. *Corpus Diacrónico y Diatópico del Español de América*. www.cordiam.org.

Chircu, A. 2014. Remarques sur l'emploi des adjectifs adverbialisés en français et en roumain. *Studii şi Cercetări Lingvistice* LXV(2): 177-187.

Gerhalter, K. 2018. Paradigmas y polifuncionalidad. La diacronía *de preciso / precisamente*, *justo / justamente, exacto / exactamente* y *cabal / cabalmente*. Dissertation thesis, Karl-Franzens-Universität Graz.

Hummel, M. in print. The third way: Prepositional adverbials in the diachrony of Romance. *Romanische Forschungen*.

Hummel, M. 2014. Adjetivos adverbiales. In C. Company Company (ed.), *Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales.* México: Fondo de Cultura Económica; Univ. Nacional Autónoma de México, 613-732.

Hummel, M. 2017. Adjectives with adverbial functions in Romance. In M. Hummel & S. Valera (eds), *Adjective adverb interfaces in Romance*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 13-46.

Hummel, M. & Gazdik, A. in preparation. *Dictionnaire historique de l'adjectif-adverbe*.

Hummel, M. & Valera, S. (eds) 2017. *Adjective adverb interfaces in Romance*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Instituto de Investigación Rafael Lapesa de la Real Academia Española. *Corpus del Nuevo diccionario histórico*. http://web.frl.es/CNDHE.

Pollin, C., Schneider, G., Gerhalter, K., Hummel, M. 2018. Semantic Annotation in the Project "Open Access Database 'Adjective-Adverb Interfaces' in Romance". *Proceedings of the Workshop on Annotation in Digital Humanities. CEUR Workshop Proceedings*. 41-46. http://ceur-ws.org/Vol-2155/.