Corpus Pragmatics
https://doi.org/10.1007/s41701-025-00199-0

ORIGINAL PAPER

Corpus Annotation of Functional Discourse Units for Aspect-Based Sentiment Analysis

Antonio Moreno-Ortiz[1] · **María García-Gámez**[1]

Received: 14 March 2025 / Accepted: 19 June 2025

Abstract

Aspect-based sentiment analysis (ABSA) aims to identify the sentiment associated with specific aspects or entities in a text. In order to facilitate the development and evaluation of ABSA systems, it is crucial to have annotated datasets that contain information about the aspects, entities, and the sentiments expressed towards them. However, the amount of information in existing datasets (for example those used in the SemEval shared tasks) is very limited. We innovate on existing corpora by introducing a multi-layered annotation schema that includes not only entities and aspects, but also lexical items and, crucially, functional discourse units (FDUs). These FDUs are text segments (typically sentences or clauses) that play a specific role or function within the overall text, such as "description", "evaluation", or "advice", a type of information which we believe can be of great help in ABSA. Our corpus focuses on user reviews of tourist attractions (specifically monuments) in the region of Andalusia (Spain), but the same schema can be used to annotate reviews of other domains simply by adapting the *aspects* layer, which is domain-dependent. The annotation schema is described, and the validation process is carried out on a sample of 400 reviews from this domain. Results show a substantial level of agreement among the annotators, indicating that the schema is reliable and consistent. We go on to illustrate and discuss some difficult cases where annotation showed discrepancy among annotators. The annotation of FDUs in the corpus is a significant advancement for aspect-based sentiment analysis.

Keywords Aspect-based sentiment analysis · Corpus annotation · Annotation schema · Functional discourse units

✉ Antonio Moreno-Ortiz
amo@uma.es

María García-Gámez
mggamez@uma.es

[1] Department of English, French and German Philology, University of Málaga, Málaga, Spain

Published online: 08 July 2025

Introduction

Opinions are influential in shaping our beliefs and behaviour according to how others view the world. When making decisions, we often seek out other people's perspectives and ask for their guidance, relying on their thoughts. Therefore, we focus on others' *evaluation*, a concept that encompasses the expression of the speaker's attitude or stance towards the entities or propositions being discussed (Thompson & Hunston, 2000, p. 5). Evaluation thus allows us to share our own ideas and compare them with those of others, often in the form of linguistic expressions, which constitute what is known as evaluative language (Benamara et al., 2017).

From a linguistic perspective, evaluation is an important area of study due to the various functions it serves, namely: (i) expressing the speaker's opinions and values; (ii) establishing relationships between the speaker and the listener; and (iii) organizing the discourse (Thompson & Hunston, 2000). The first function involves sharing the speaker's ideology, implicitly or explicitly. The second uses evaluative language to signal degrees of certainty or commitment—subsuming hedges, boosters, and other stance markers—without conflating these devices with deliberate truth-value manipulation (Hyland, 2005). The third function goes beyond the speaker-listener dyad to create textual coherence: by deploying evaluative cues, the author guides readers through the argument, signaling introductions, transitions, and conclusions (Thompson & Hunston, 2000). While the first two functions have been widely discussed in the literature on evaluative language, particularly by Appraisal Theory (Martin & White, 2005), the third one has been largely ignored in computational treatments of evaluative language due to processing difficulties (Benamara et al., 2017). This piece of research seeks to dive deeper into the possibilities that this third function of evaluative language has to offer. In particular, we focus on how aspect-based sentiment analysis, which we define in the next section, can benefit from the availability of annotated corpora at the discourse level with a special emphasis on discourse functions. While our research has a clear practical application in this field, it also extends and validates previous studies that use this discourse analysis framework, such as Biber et al. (2007), Vásquez (2011), and Egbert et al. (2021).

Aspect-Based Sentiment Analysis

Sentiment analysis is inherently concerned with evaluative language, as it is a field of study that analyzes "people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as product, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2011, p. 459). Its basic tasks are polarity detection and emotion recognition, which is typically implemented as a classification task (Cambria et al., 2017), although it may also attempt to reach a more fine-grained level of language understanding by trying to identify "not only the overall semantic orientation of a text, but also the specific,

possibly phrase-level evaluative utterances, as well as the aspects of the entities they refer to" (Moreno-Ortiz et al., 2019, p. 2), which is known as aspect-based sentiment analysis (ABSA henceforth).

The tools and techniques used in sentiment analysis can employ machine learning models and algorithms, sentiment dictionaries, or a combination of both. In the case of the machine-learning approach, it uses a set of features which are learned from annotated corpora or labeled examples. Although machine learning approaches have reached high levels of performance in language data analysis, they still perform poorly in domains they have not been previously trained on (Aue & Gamon, 2005; De Clercq et al., 2017; Wang & Liu, 2015). The lexicon-based approach, on the other hand, uses a dictionary to provide the polarity for each word or phrase found in the text, sometimes in combination with some system to account for the immediate context, i.e., the so-called *valence shifters* (Muhammad et al., 2016; Taboada, 2016). This one also presents some drawbacks, as it requires rich lexical knowledge to achieve good results in different domains (Moreno-Ortiz & Pérez-Hernandez, 2018). Moreover, contextual valence shifters present a challenge both at the sentence and discourse levels: for instance, in terms of higher-order linguistic levels of analysis, phenomena such as metaphors, sarcasm, understatements, or humblebragging can determine shifts in polarity. This is especially true of social media, one of the major genres that sentiment analysis has been applied to; yet, no practical solutions have been offered for this problem yet (Moreno-Ortiz, 2024).

ABSA also poses a challenge for Natural Language Processing (NLP) because, as expressed by Moreno-Ortiz et al. (2019), it requires the existence of annotated corpora "on which to train classification algorithms and/or learn the concepts and lexical items involved in the evaluative texts" (p. 3), and requires the design of an appropriate annotation schema. Nevertheless, the literature on schema generation and validation for ABSA is scarce, and available annotated corpora are few and follow a simplistic approach, as exemplified by the SemEval datasets (Pontiki et al., 2014, 2015, 2016), which also focus on the hospitality industry (lodging and catering in their case, tourist attractions in ours). These annotated corpora are limited to one layer that identifies the aspect and orientation of the evaluative expressions, as well as the entities they refer to. While these resources have been successfully used to train machine learning models that produce reasonable results, we think there is ample room for improvement, as higher quality and accuracy could be achieved by incorporating discourse-level information in the form of functional discourse units (FDUs henceforth) and by clearly separating the rest of the annotations in different layers.

Corpus Annotation for Discourse Analysis

Corpus annotation at the discourse level has traditionally centered on the study of *discourse relations*. This focus stems from a fundamental assumption in discourse analysis: texts are not a random sequence of sentences; rather, the order of those sentences matter, as do the relations that those sentences hold to one another (e.g., cause and effect, contrast, reason, explanation, illustration). The significance of

these relations is the existence of an entire class of grammatical words—connectors, primarily conjunctions and adverbs—whose sole function is to signal them.

From a discourse analysis perspective, connectors are often studied within the broader category of discourse markers, which encompasses all linguistic elements that primarily serve a pragmatic role in structuring a text and managing the flow of discourse. While connectors fulfill grammatical functions such as coordination ("and," "but," "or") or subordination ("although," "because"), discourse markers typically lack syntactic functions and are often classified as "disjuncts," meaning they are syntactically independent elements. Rather than contributing to sentence structure, they signal the speaker's attitudes, guide the reader's or listener's expectations, and shape discourse interpretation. Discourse markers can be organizational ("firstly," "in essence," "finally"), conversational ("I mean," "you know"), or modal and hedging markers ("in general," "perhaps," "sort of").

Given their crucial role in text cohesion and, more specifically, coherence, connectors and discourse markers have been widely studied in discourse analysis. Coherence, one of the central concerns of the field, is defined by Van Dijk (1977) as "(...) a semantic property of discourse, based on the interpretation of each individual sentence relative to the interpretation of other sentences" (p. 96).

Discourse connectors and markers are among the most salient elements for identifying coherence, as they provide the structural framework that organizes and unifies a text thematically. Comprehensive inventories of these elements have been developed; for instance, Kalajahi et al. (2017) compiled a list of 632 discourse connectors, categorized into eight broad classes and 17 subcategories. Consequently, corpus annotation initiatives in discourse analysis have also predominantly focused on discourse connectors.

One of the most relevant annotation projects in this field, the Penn Discourse Treebank (PDTB) (Prasad et al., 2008, 2018), initially adopted a "lexically grounded" approach, in which discourse relations were identified exclusively through explicit discourse connectors (referred to as "connectives" in their framework). However, discourse relations are not always overtly marked; speakers and writers frequently rely on the audience's inferential abilities to establish connections between textual elements. For example, in the statement "I thought it was pricey. I decided not to buy it." there is no explicit connective, yet the causal relationship between the two sentences—where the decision not to purchase is attributed to the high price—is easily inferred through adjacency and general knowledge.

Recognizing this limitation, the third version of the Penn Discourse Treebank (PDTB-3) expands its annotation framework to include both implicit discourse relations, which must be inferred, and "NoRel" cases, where adjacent sentences bear no discernible relation. This refinement acknowledges the complexity of coherence, moving beyond explicit markers to capture the full range of discourse relationships.

This consideration is particularly relevant for social media texts, which often lack coherence and may consist of loosely related or even disconnected statements. Given this characteristic, we opted for a functions-based approach to annotation—described in detail below—rather than the more conventional, relations-based approach commonly used in discourse-level corpus annotation. This

decision allows for a more flexible and context-sensitive analysis, better suited to the fragmented and dynamic nature of social media discourse.

Discourse relations are also referred to as "rhetorical relations," a term rooted in a specific theoretical approach to discourse analysis: Rhetorical Structure Theory (RST). RST, a linguistically motivated framework developed by Mann and Thompson (1987, 1988), has significantly contributed to the development of discourse-level annotated resources. This theory adopts a descriptive approach aimed at identifying textual structure by segmenting discourse into elementary discourse units (EDUs). These EDUs are interrelated, reflecting the organization and coherence of a text while forming a hierarchical structure.

RST is based on the premise that every segment of a text serves a distinct function within an overarching hierarchical organization, where higher-order structures are composed of more fundamental EDUs. Within this framework, two primary types of units are distinguished: (a) the nucleus, which conveys the core meaning of a passage, and (b) satellites, which provide supplementary information. This distinction underscores the functional asymmetry between discourse units, reinforcing the hierarchical nature of textual organization.

RST has proven valuable in lexicon-based sentiment analysis tasks, as demonstrated by Voll and Taboada (2007) and Taboada et al. (2008), whose pioneering work introduced discourse information into sentiment analysis to determine overall document polarity—an essential requirement for lexicon-based approaches. More recently, Huber and Carenini (2020) explored discourse-augmented sentiment analysis by leveraging discourse tree structures combined with hierarchical neural architectures.

The underlying rationale is that incorporating discourse information allows sentiment-laden lexical units to be contextualized, enabling their valence to be adjusted based on the type of text segment EDU in which they appear. This approach highlights the importance of discourse relations in identifying EDUs and enhancing the interpretive weight of sentiment-laden expressions. Our methodology is similarly grounded in this principle.

However, RST does not always perform as expected. As reported by Moreno-Ortiz (2024), RST parsers exhibit low reliability in research for two main reasons: (i) they depend on syntactic parsers whose performance is often suboptimal, particularly with social media texts—our primary focus—where misspellings, unconventional syntax, and inconsistent or absent punctuation are common; and (ii) different parsers frequently yield divergent analyses of the same text, with no clear indication of which parse is the most accurate. This issue stems from RST's original design for formal text genres, typically written by specialists who adhere to established linguistic norms in terms of spelling, punctuation, grammar, and textual organization—features that are often lacking in online user-generated content. Similarly, Ein-Dor et al. (2022) argue that traditional discourse-based sentiment models struggle with informal text structures, particularly in user-generated content such as social media posts. To address this, they propose an alternative approach using sentiment-carrying discourse markers to generate large-scale weakly labeled training data, which is then used to fine-tune language models.

Moreover, RST parsers have demonstrated poor performance in aspect-based sentiment analysis (ABSA). Schouten and Frasincar (2016) specifically investigated the practical application of RST to ABSA—our primary application task—and found that, while their methodology generated high-quality linguistic information, its results did not match the effectiveness of machine learning-based approaches.

Given these limitations, it becomes evident that a relations-based discourse annotation framework, with its reliance on hierarchical structures and formal linguistic markers, is not optimal for analyzing the fragmented and dynamic nature of social media discourse.

Genre Analysis, Moves, and Functional Discourse Units

Our concept of functional discourse units (FDUs) is based on the work of Egbert et al. (2021), who employed a similar methodology to analyze spoken language using the *British National Corpus (BNC) Spoken 2014*. Their approach involved segmenting transcribed conversations into discourse units and classifying them according to their communicative purposes. This methodology builds upon *move analysis*, originally introduced by Swales (1981, 1990) for the analysis of research articles; *moves* are functional units within a text, each fulfilling a distinct communicative function. Although moves may vary in length, they typically contain at least one proposition. Within the updated framework (Swales, 2004), known as the *Create a Research Space (CARS) model*, Swales proposed a three-move schema for analyzing research article introductions, which consists of: (i) establishing a territory, (ii) establishing a niche, and (iii) occupying the niche.

Over the years, this function-based approach—under various appellations—has been applied to a wide range of genres, including legal discourse (Bhatia, 1983) and philanthropic discourse (Upton & Connor, 2001) to movie reviews (Pang, 2002). Notably, it has also been successfully employed in the analysis of online consumer reviews. For example, Zhang and Vásquez (2014) examined 80 hotel replies that were posted in response to online complaints, while Panseeta and Todd (2014) analyzed 100 responses to negative reviews on TripAdvisor, written by 20 Thai hotel management teams. More recently, Cenni (2024) investigated 300 negative hotel reviews posted on TripAdvisor—100 in each of three languages: Spanish, Italian, and French.

Building on this tradition, this article introduces a novel approach for fine-tuning pre-trained language models using discourse information extracted from annotated corpora. Our primary objective is to develop a robust annotation schema for aspect-based sentiment analysis (ABSA) with discourse features, a key goal of our DisParSA project. This schema is grounded in the concept of *discourse units*, as formulated by Biber et al. (2007), which we consider a more suitable alternative to Rhetorical Structure Theory (RST) for our purposes. While this concept aligns with RST's core assumption that every segment of text serves a communicative function, it avoids constraining these functions to the clause level or mandating explicit relational links between units. Instead, it prioritizes the *functional roles* that segments

play within a text, rather than their interrelationships, offering a more flexible and context-sensitive approach to discourse annotation.

Methods and Corpus

The primary objective of this work is to design, implement, and validate an annotation schema for aspect-based sentiment analysis with rich discursive features, and which includes four layers of discourse annotation: FDUs, aspects, lexicon, and entities (described in section Annotation Schema). This work is part of a bigger project, entitled DisParSA, which has two main objectives: (i) to provide information on how evaluation is linguistically expressed (i.e., how speakers structure the information when they state their opinions), and (ii) to use the data to finetune a pretrained Transformer model so that it can be used to identify these segments automatically. Designing a valid annotation schema is therefore crucial to the development of the project, as is the case with any large-scale corpus annotation effort. In order to develop this schema we follow the procedure described by Moreno-Ortiz et al. (2019), whose work closely aligns with our research, as it also focuses on aspect-based sentiment analysis of user-generated hotel reviews. Thus, the annotation schema was developed following a systematic procedure:

1. Definition of attributes and values for each annotation layer. Each layer presents distinct challenges, with some being more complex and less tested than others.
2. Preliminary annotation of a small sample to informally validate the schema. This early step helps identify potential issues before conducting a full validation via multi-annotator agreement.
3. Annotation of a substantial sample (400 reviews) by three independent annotators, followed by agreement analysis. Since our annotation tool of choice, *Prodigy*, does not natively support multi-layer annotation, we developed a custom Python script to handle this process. Agreement metrics for each annotation layer were computed using the *Disagree* Python package.[1]
4. Exploratory analysis of annotation results, including statistical evaluation and visualization. A separate Python script, using the *Altair* graphics library, was developed to compute relevant statistics and generate visual representations of the annotated data.

The corpus consists of English-language user reviews of monuments in Andalusia, Spain, sourced from TripAdvisor and Google Reviews. The dataset encompasses a wide variety of cultural heritage sites, including statues, churches, viewpoints, and historical landmarks. In total, the corpus comprises 7,145 user reviews, amounting to approximately 678,000 words. For the annotation process, a random sample of 400 reviews was selected, applying a minimum length filter of 30 words

[1] https://github.com/o-P-o/disagree

to ensure sufficient textual content for discourse analysis. These reviews were then annotated by three human annotators following the schema and workflow detailed in the next section. The annotated sample comprises over 38,000 words, with an average of 90 words per review and a range of 30 to 194 words.

As Cenni (2024), observes, online travel reviews represent the most pervasive written genre through which tourists share their travel experiences online (p. 77). These spontaneous and publicly accessible user-generated texts not only reflect personal experiences but also provide valuable guidance and cultural insights for prospective visitors. Consequently, online reviews serve both as expressions of subjective evaluation and as informational resources, making them a particularly relevant genre for aspect-based sentiment analysis with discourse-driven annotation.

Annotation Schema

The proposed annotation schema has, as mentioned above, four annotation layers, one for each of the categories considered relevant for aspect-based sentiment analysis. These are aspects (the relevant aspects that are found in the evaluative expressions in user reviews), entities (the types of entities mentioned in reviews, the main one being the evaluated entity, i.e., the monument), lexical items (the words and phrases that denote evaluation), and discourse functions.

This schema attempts to go a step further than previous ABSA annotation attempts in several ways. The lexical layer (sentiment words and expressions) is commonly found in the literature (Pontiki et al., 2014, 2015, 2016), as is the set of entities, although we not only annotate the evaluated entities, but other types as well (see below); this is necessary because other entities (e.g., opinion holders other than the reviewer, other evaluated entities) are commonly mentioned in the texts that need to be identified. The most original addition is the discourse functions layers (FDUs). As we have seen, user reviews have indeed been analyzed from this perspective (Cenni, 2024; Vásquez, 2011; Zhang & Vásquez, 2014) but, to our knowledge, no formal annotation projects have been carried out before.

Aspects

The first step in an ABSA-oriented annotation effort, and probably the hardest, is to identify the specific aspects that are evaluated in a given domain. Aspects are domain-specific, as different characteristics are liable to be evaluated. For example, in the domain of hotels, users will comment on aspects like price, staff, beds, toiletries, or amenities, whereas in a monument, only some of these aspects are equally relevant (e.g., price), others may be present but are less common (e.g., staff), and others are not present at all.

Although some of these aspects can be guessed, our approach was strictly data driven, using the contents of the corpus as the only guide to discover the set of relevant aspects. Thus, we used the *Sketch Engine* (Kilgarriff et al., 2014) corpus query suite to identify the set of aspects relevant to the domain of user reviews of

Table 1 List of aspects considered by the annotation schema

Label	Description
Accessibility	Access to the monument.
Artistic value	Perceived value of works of art present.
Atmosphere	Perceived atmosphere at the site.
Crowds	Comments on the number of visitors and waiting times.
Facilities	Additional parts of the monument, such as cafés, bathrooms, cloakrooms, etc.
General	General comments (e.g., "it wasn't worth it").
Info	Quality of information panels or other devices about the monument.
Location	Geographical situation of the monument.
Maintenance	Conservation of the monument.
Price	Cost of the ticket to enter the monument.
Safety	Perceived personal safety/security at the site and surroundings.
Size	Comments on the size of the monument.
Staff	Personnel working at the monument.
Views	Views from the monument.

monuments by searching the corpus for prototypically positive and negative verbs, deverbal nouns (and looking at their objects), and qualifying adjectives (and looking at the nouns they modify). The Word Sketch tool is particularly useful to achieve this, as it uses the built-in part-of-speech tagging and syntactic parsing capabilities of Sketch Engine for user-provided corpora to summarize results in a very user-friendly way. Thus, looking up prototypically positive and negative adjectives, such as "good" or "terrible" in the Word Sketch returns a very useful snapshot of how those words are used in the corpus, including what nouns they modify, which gives us very accurate pointers to the information we require, i.e., what types of entities are being evaluated. Moreover, we performed lookups of verbs such as "like", "enjoy", or "love" and checked which were the objects of these verbs. We also searched for nouns that denote domain-specific entities (e.g., "monument", "landmark", "church") and looked at their modifiers. For instance, the aspect *size* was evidently relevant, as many monuments were modified by adjectives such as "big", "small", or "huge". On the other hand, terms such as "professional", "efficient", or "fast" (to name a few) hinted at the need for a *staff* aspect.

After comparing this process with further manual corpus browsing via concordances—and, above all, insights from our preliminary annotation trial and the annotators' feedback session (see Section"Schema Validation")—we identified eleven aspects that tourists commonly use to evaluate monuments. Table 1 shows this list in alphabetical order and a description of each aspect. We discuss the relevance — determined by frequency— of each aspect in Section"Results".

Entities

Identifying the types of entities that are mentioned in the user reviews is necessary in order to know what exactly is being evaluated. This is because not all comments

Table 2 List of entity types

Label	Description
Evaluated entity	The main entity under evaluation.
Evaluated entity part	Particular sections or components of the main entity.
Other evaluated entity	An external entity under evaluation, often by comparison with the main evaluated entity.
Opinion holder	Writer of the review, the person who is telling their opinion.
Other opinion holder	A person other than the writer of the review and whose opinion is included in the text.
Reader	The person who is reading the review.

present in reviews are directed at the evaluated entity itself, some of its aspects, or even its parts. Instead, users may include comparisons with past experiences, which may also be evaluated. They might also mention the target audience and/or opinion holders other than themselves (e.g., friends or family members that have also visited a monument). Although this task is easier than identifying the aspects, we decided to be systematic. Therefore, the procedure included the creation of a simple Python script that used Spacy to identify entities and then manually categorizing the extracted list by looking at the contexts in which they appeared. The final list of types of entities mentioned by users in monument reviews (and possibly in any other type of user review) is shown in Table 2.

Sentiment Lexical Items

Annotation of lexical items with semantic orientation is key both for lexicon-based sentiment analysis, where this is the only feature that is employed by the analyzers, and in machine learning-based classifiers, where the use of sentiment dictionaries, although strictly necessary, has been shown repeatedly in the literature to improve results (Gaikwad & Joshi, 2016; Kolchyna et al., 2015).

In an attempt to replicate the analysis of a lexicon-based analyzer, we opted for a domain-independent annotation of sentiment-laden lexical units. Specifically, only those words and expressions that inherently convey semantic orientation, irrespective of subject matter, were annotated. In sentiment analysis, domain specificity plays a crucial role in determining the polarity of many expressions. This is one of the primary reasons why machine-learning approaches have traditionally outperformed lexicon-based models, as the former are typically trained on domain-specific datasets. For instance, in the hospitality industry, location is among the most frequently mentioned aspects in user reviews. Consequently, expressions such as "centrally located" or "restaurants and shops nearby" are perceived as positive within this domain, despite not containing any explicit sentiment words.

Although a number of lexicon-based systems have attempted to incorporate domain-specificity—e.g., Shaukat et al. (2020), Muhammad et al. (2016), Moreno-Ortiz (2017)—most available lexicons and systems do not consider this key feature. Our annotation approach, however, considers full sentiment expressions, including

Table 3 List of sentiment lexical items types

Label	Description
Positive lexical items	Words or phrases with a positive semantic orientation.
Negative lexical items	Words or phrases with a negative semantic orientation.

valence shifters—words or phrases that invert or modify sentiment polarity. For example, expressions like "extremely expensive" or "not worth it" alter the overall sentiment conveyed, requiring their explicit identification in annotation. Therefore, our sentiment lexicon annotation layer follows this simplified approach. Table 3 lists the two labels considered in this layer, which is limited to (domain-independent) positive and negative lexical items.

Functional Discourse Units

This annotation layer constitutes the key distinguishing feature of our proposal. It is grounded in the theoretical framework outlined in 5.3 above, which posits that text segments fulfilling specific communicative functions within a given genre—user reviews, in this case—should be explicitly labeled as such. Our annotation guidelines adhere to the premise that every section in a text fulfils a particular function, and therefore every section of a text should be labeled for FDUs. The length of these segments is determined by the functions they perform, with a general tendency for FDUs to be realized through simple sentences or clauses within complex sentences.

Determining the precise boundaries of each text segment can be challenging, particularly in user reviews and social media texts, which often deviate from standard writing conventions in terms of spelling, grammar, and punctuation. This variability calls for a "relaxed" approach to inter-annotator agreement calculation, as different annotators may select slightly different text spans (in terms of character length) while still referring to essentially the same discourse unit. A more detailed description of this approach is provided in Section"Schema Validation".

Regarding the methodology used to identify FDUs, we initially adopted the move taxonomy proposed by Cenni (2024), formed by three moves: (i) extra background information, (ii) evaluations (negative or positive), and (iii) future-oriented recommendations. To enhance clarity and efficiency, we shortened these labels, renaming them as *context*, *positive evaluation*, *negative evaluation*, and *advice*. Nevertheless, after examining the corpus we noticed that review authors often provided factual information about the monument, and that these text spans did not fall under any of the above-mentioned labels. Consequently, we introduced a fifth category: *description*. Table 4 provides an overview of the final set of FDUs along with an explanation of each of category.

Table 4 List of FDU types

Label	Description
Advice	Practical recommendations about the visit (e.g., optimal times, saving options)
Context	The writer puts their visit into context and may provide personal details.
Description	Objective, factual information about the monument.
Positive evaluation	Positive comments about the monument.
Negative evaluation	Negative comments about the monument.

Table 5 Final annotation schema

Layer	Label
FDU	ADVICE CONTEXT DESCRIPTION EVAL_POS EVAL_NEG
ASP	**ARTISTIC_VALUE** ACCESSIBILITY **ATMOSPHERE** **CROWDS** FACILITIES GENERAL INFO LOCATION MAINTENANCE PRICE SAFETY SIZE STAFF VIEWS
LEX	LEX_POS LEX_NEG
ENT	OP_HOLDER OP_HOLDER_OTHER EVAL_ENTITY **EVAL_ENTITY_PART** OTHER_EVAL_ENTITY READER

Categories marked in bold typeface were not part of the initial set prior to the feedback session with the annotators, which was part of the validation process, as described in the following section.

Final Annotation Schema

The final annotation schema comprises these four annotation layers: functional discourse units (FDU), aspects (ASP), lexicon (LEX), and entities (ENT). Table 5 summarizes the overall schema with the actual labels we use in annotation, which, by

convention, are formed by a combination of three letters that define the layer, followed by a colon and a tag that defines the type or class (e.g., "FDU: ADVICE").

Schema Validation

The annotation schema is governed by a set of rules designed to ensure consistency across annotators throughout the annotation process. First, all evaluative FDUs must include at least one aspect, with a maximum of two. These aspects must correspond to the text spans covered by the evaluative FDUs. In contrast, non-evaluative FDUs may include aspects, but their presence is optional (e.g., a description may reference a particular aspect, but it is not required to do so). Additionally, contextual modifiers must be annotated to capture the full scope of sentiment expressions. For instance, in "absolutely beautiful," the modifier "absolutely" must be included in the annotation.

To enhance reliability and reduce annotation inconsistencies across annotators, they were instructed to follow a specific sequence in the mark-up procedure:

1. Identification of FDUs within the context.
2. Annotation of aspects contained within evaluative FDUs.
3. Marking of sentiment-laden lexical items found in the review. This step was designed to prevent cognitive biases, such as the assumption that the presence of an evaluative word necessarily implies an evaluative function.
4. Identification of entities mentioned in the text.

All reviews were annotated by three human annotators using Prodigy (Montani et al., 2021), a server-based corpus annotation tool designed for machine learning purposes. Since Prodigy does not support multilayer annotation natively, a naming protocol was devised where all layers conflate in one and the first three letters of each label define the layer (FDU, ASP, LEX, ENT).

To validate the annotation schema, two trials of fifty texts each were scheduled prior to tackling the full set of reviews. Annotators were asked to identify and mark the text segments and keep a record of the issues that they encountered. After the first trial, we calculated both Krippendorff's *alpha* (α) (Krippendorff, 2004) and Fleiss's *kappa* (κ) coefficients (Fleiss, 1981), which are suitable to three annotators or more. Krippendorff's *alpha* (α) informs about the reliability of the data (i.e., whether the results can be reproduced), and Fleiss's *kappa*, being an extension Cohen's Kappa for more than two annotators, provides a score of the validity of the data on which the annotators agree (Artstein & Poesio, 2008). Their values range from 0 (chance agreement) to 1 (full agreement). As mentioned above, the Python Disagree package was used to calculate both metrics.

For two annotations to be considered a match, both the annotated text span and the assigned label must be identical. While label matching is straightforward to assess, text span matching presents additional challenges, particularly in cases where spans partially overlap (i.e., when two annotations cover the same content but differ slightly in length). To address this, we computed the overlap ratio, which represents

Table 6 Inter-annotator agreement metrics for the first trial

Layer	Krippendorf's *alpha* (α)	Fleiss's *kappa* (κ)	Interpretation
FDU	0.598	0.601	Moderate agreement
ASP	0.537	0.538	Moderate agreement
LEX	0.698	0.700	Substantial agreement
ENT	0.521	0.521	Moderate agreement

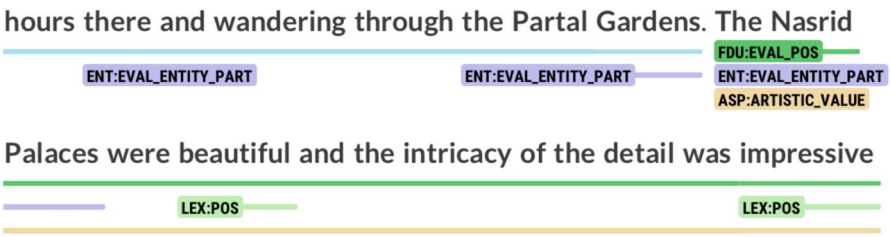

Fig. 1 Example of the aspect ARTISTIC_VALUE

the percentage of shared characters or tokens between two annotations. A minimum threshold of 70% character overlap was established to determine whether two annotations could be considered a match. The results of these agreement metrics for each annotation layer are summarized in Table 6.

Given the agreement policy and the complexity of the annotation task, a relatively low inter-annotator agreement (IAA) was anticipated. Substantial agreement was achieved for the lexicon layer, whereas the lowest agreement scores were observed in the entities layer. This discrepancy can be attributed to misinterpretations regarding what needed to be annotated—some annotators omitted parts of the main entity or first-person pronouns referring to the reviewer (opinion holders). To address these issues, a feedback session was conducted in which annotators discussed their concerns, particularly regarding the annotation of FDUs and aspects. The primary source of confusion stemmed from unclear syntactic guidelines—annotators were initially instructed to mark "spans" without a predefined syntactic unit (e.g., clauses or sentences). To resolve this ambiguity, the annotation instructions were refined, explicitly directing annotators to annotate FDUs and aspects at the clause level. Additionally, annotators suggested incorporating three additional aspects that were not covered in the initial annotation schema: *artistic_value* (paintings, sculptures, and other artistic elements present at the monument), *atmosphere* (tone or mood at the site and its surroundings), and *crowds* (amount of people visiting the monument and waiting times). Figures 1, 2, and 3 show examples of these aspects.

For the entities layer, annotators observed that the original schema only accounted for the evaluated entity as a whole, without considering its individual parts. This issue was particularly recurrent in reviews of monuments such as churches, where elements like altars, chapels, and side aisles were frequently evaluated independently. To address this limitation, we introduced the *eval_entity_part* label. Fig. 4 presents an excerpt from a

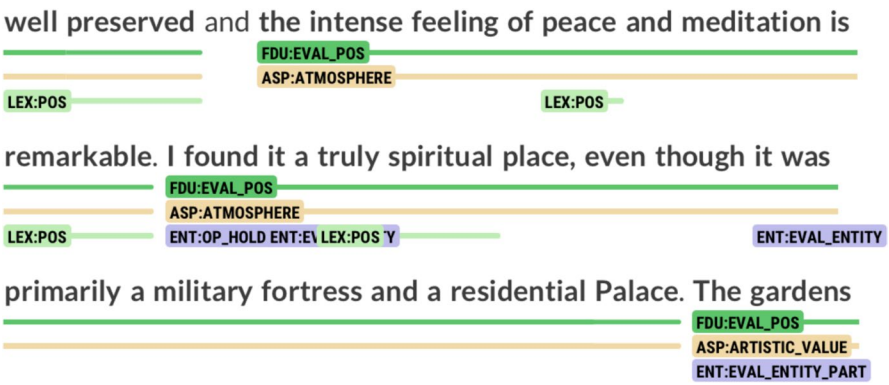

Fig. 2 Example of the aspect ATMOSPHERE

Fig. 3 Example of the aspect CROWDS

review in which gardens are evaluated as an entity. Under the original annotation schema, these gardens would have either been annotated as *eval_entity*—despite not being the primary entity under evaluation—or omitted entirely.

After achieving consensus and incorporating these modifications to the schema, a second trial of 50 annotations was carried out. To validate the updated schema, we recalculated Krippendorff's alpha (α) and Fleiss's kappa (κ) coefficients to assess IAA (Table 7).

The improved results show the importance of feedback sessions during the schema design phase. Although the annotation of FDUs and the aspects still presented a relatively high degree of disagreement among annotators, this was an expected outcome, as these layers are more open to interpretation by nature.

Results

Finally, annotators proceeded with the annotation of a sample of 400 reviews following the annotation schema previously described in Section Annotation schema. To determine which annotations would be included in the final annotated corpus, a majority vote rule was applied: if two annotators agreed on a given annotation while the third deviated, the

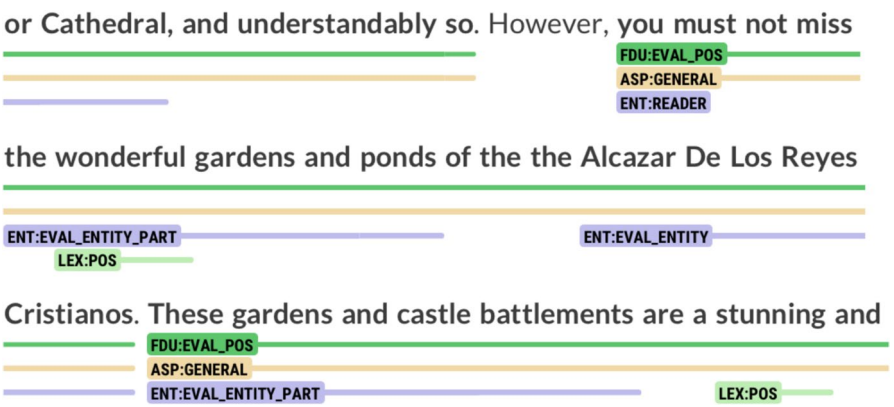

Fig. 4 Annotation of EVAL_ENTITY_PART

Table 7 Inter-annotator agreement metrics for the second trial

Layer	Krippendorf's *alpha* (α)	Fleiss's *kappa* (κ)	Interpretation
FDU	0.680	0.681	Substantial agreement
ASP	0.671	0.671	Substantial agreement
LEX	0.697	0.698	Substantial agreement
ENT	0.712	0.712	Substantial agreement

Table 8 Inter-annotator agreement metrics for the full annotated corpus

Layer	Krippendorf's *alpha* (α)	Fleiss's *kappa* (κ)	Interpretation
FDU	0.698	0.700	Substantial agreement
ASP	0.685	0.685	Substantial agreement
LEX	0.705	0.706	Substantial agreement
ENT	0.700	0.701	Substantial agreement

consensus annotation from the majority was selected.[2] As shown in Table 8, the IAA metrics improved marginally compared to the two previous trials.

[2] The entire annotated corpus is available at https://osf.io/... [The full URL will be provided once the article review process has been completed in order to ensure anonymization].

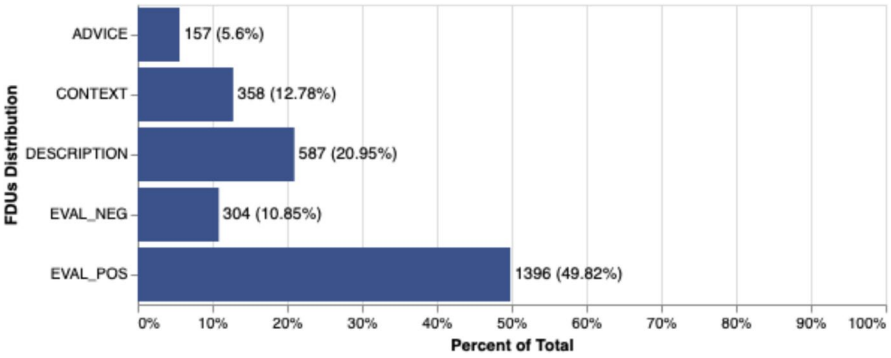

Fig. 5 Distribution of annotated FDUs

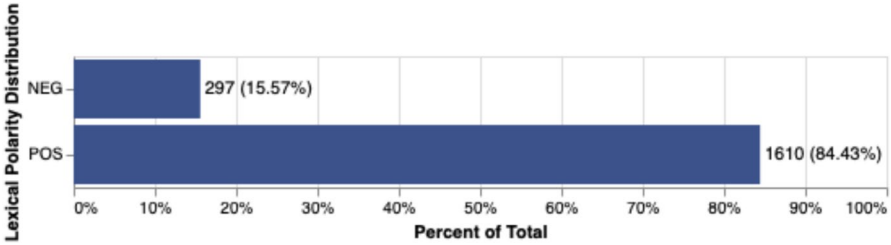

Fig. 6 Distribution of annotated lexical items

Regarding the annotated FDUs, Fig. 5 presents the distribution statistics.[3] As shown, *positive evaluations* were the most prevalent function, accounting for 49.82% of all annotated text segments. This category was followed by *description* (20.95%), a function that has not been explicitly recognized in previous research but appears to play a significant role. *Negative evaluations* comprised only 10.85% of the sample, while *advice* accounted for 5.6%.

These findings suggest that users generally emphasize the positive aspects of monuments in their reviews while downplaying negative ones. While the prevalence of positive comments in user reviews has been widely discussed in the literature (Panseeta & Todd, 2014; Vásquez, 2011; Zhang & Vásquez, 2014), our study offers empirical evidence derived from a systematically annotated, sizable dataset. Additionally, these results highlight that objective, descriptive comments are more common than previously assumed.

Regarding lexical items, results correlate with the high proportion of positive evaluation utterances in the sample, shown in Fig. 6. The correlation is confirmed by the heatmap in Fig. 7.

[3] A Python script was created to extract and process annotations from the corpus. This script generates lists of annotations for each layer in HTML format and generates the data visualizations that we use in this report.

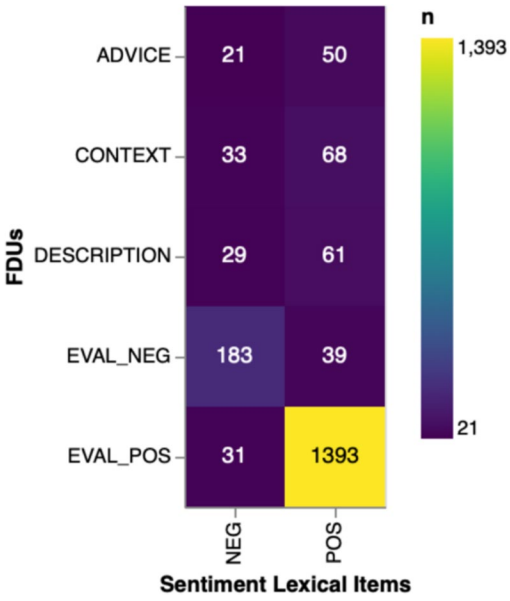

Fig. 7 Correlation heatmap between FDUs and sentiment-laden lexical items

Regarding the annotation of aspects, Fig. 8 presents the distribution of each category within the sample. As shown, *general* is the most frequently annotated aspect, appearing in 37.82% of cases, followed by *artistic_value* (15.81%). The remaining aspects each account for less than 10% of the instances in the corpus. Notably, *safety* was the least annotated aspect, appearing in just 0.27% of cases. While the COVID-19 pandemic has officially ended and restrictions have been lifted, we still consider safety concerns to be relevant in the current context, which is why this aspect was included in our annotation schema.

The results for entity annotation are presented in Fig. 9, and are in accordance with expectations: *eval_entity* accounts for most instances (40.98%) whereas *eval_entity_part* appear in only 13.25% of the sample. Appeals to readers are relatively frequent, comprising 16% of the annotated instances. In contrast, references to other evaluated entities are rare, appearing in only 4.29% of the sample, which suggests that users seldom engage in comparative evaluations when reviewing monuments.

To finish this section on the quantitative results, Table 9 offers a summary of all layers and categories, along with an example extracted from the corpus for each of them.

Discussion

Despite the high agreement rate among annotators, certain dissonances emerged in specific cases. One notable source of disagreement involved FDUs such as *context* and *eval_pos*, which were occasionally confounded by annotators. Figure 10 presents an excerpt from a review describing a visit to Santa María La Mayor Coronada,

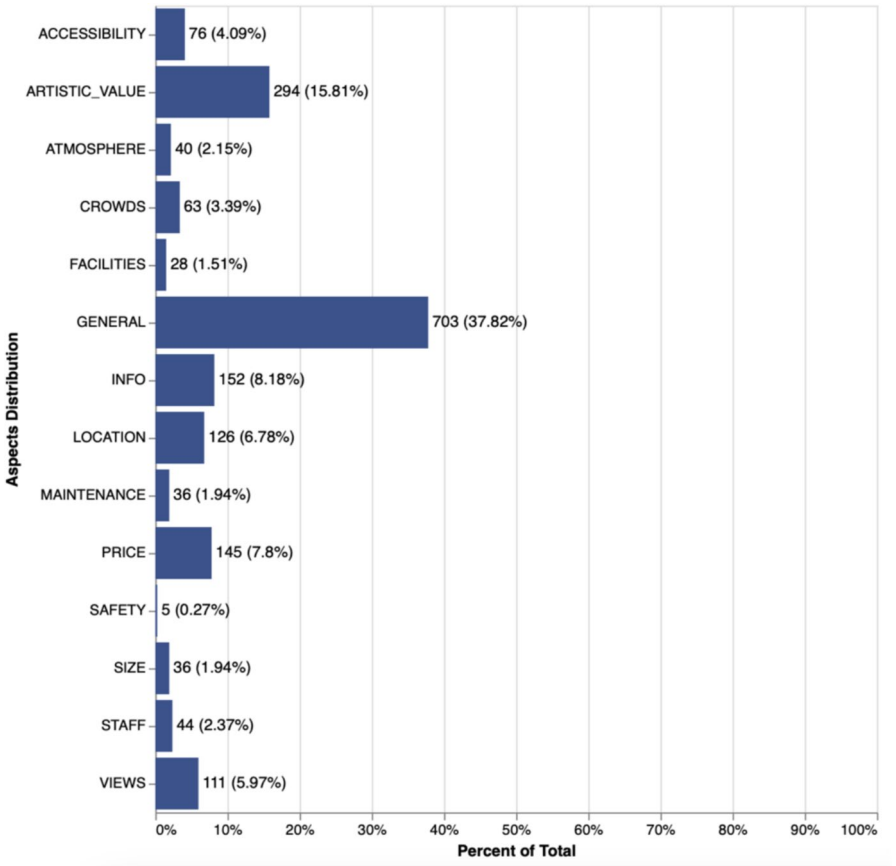

Fig. 8 Distribution of aspects

a church located in Medina Sidonia, Cádiz. Annotator A classified this segment

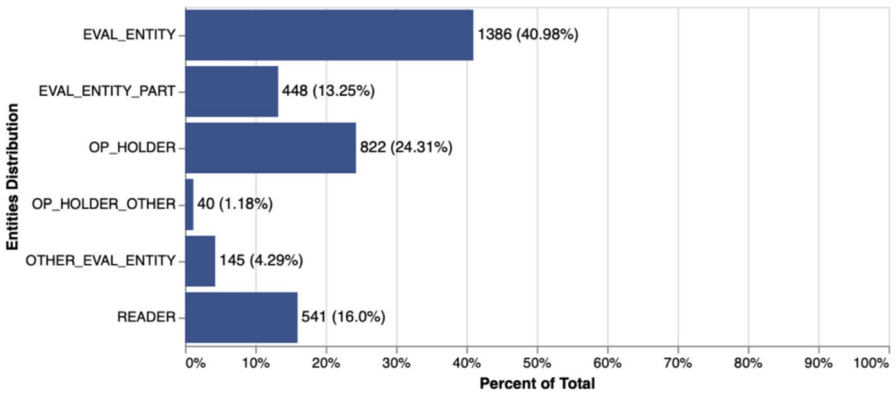

Fig. 9 Distribution of entities

Table 9 Layers and labels of the annotation schema sorted by frequency

Annotation layer	Label	Frequency	Example
FDU	EVAL_POS	49.82%	"it was definitely worth it"
	DESCRIPTION	20.95%	"it was built in 1918"
	CONTEXT	12.78%	"we had always wanted to visit this monument"
	EVAL_NEG	10.85%	"this place was horrible"
	ADVICE	5.6%	"you should queue early"
ASP	GENERAL	37.82%	"we loved it"
	ARTISTIC VALUE	15.81%	"beautiful art"
	INFO	8.18%	"audio guides are provided for free"
	PRICE	7.8%	"the site offered poor value for money"
	LOCATION	6.78%	"it is located in the city centre"
	VIEWS	5.97%	"the views from the roof were breathtaking"
	ACCESSIBILITY	4.09%	"it has a ramp for wheelchair users"
	CROWDS	3.39%	"it was incredibly crowded"
	STAFF	2.37%	"the guide at the museum was very nice"
	ATMOSPHERE	2.15%	"this was a truly spiritual place"
	SIZE	1.94%	"the museum is very small"
	MAINTENANCE	1.94%	"the walls needed some painting"
	FACILITIES	1.51%	
	SAFETY	0.27%	"there were some pickpockets"
LEX	LEX_POS	84.43%	"beautiful place"
	LEX_NEG	15.57%	"very ugly space"
ENT	EVAL_ENTITY	40.98%	"the monument was beautiful"
	OP_HOLDER	24.31%	"we visited this church"
	READER	16%	"you must come here"
	EVAL_ENTITY_PART	13.25%	"the altar of the church"
	OP_HOLDER_OTHER	4.29%	"some people say that this church is not worth a visit"
	OTHER_EVAL_ENTITY	1.18%	"the alhambra was more interesting than this monument"

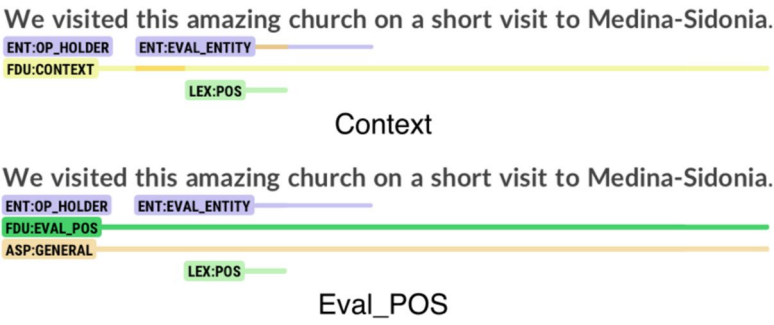

Fig. 10 Annotation of a segment as *context* or as *eval_pos*

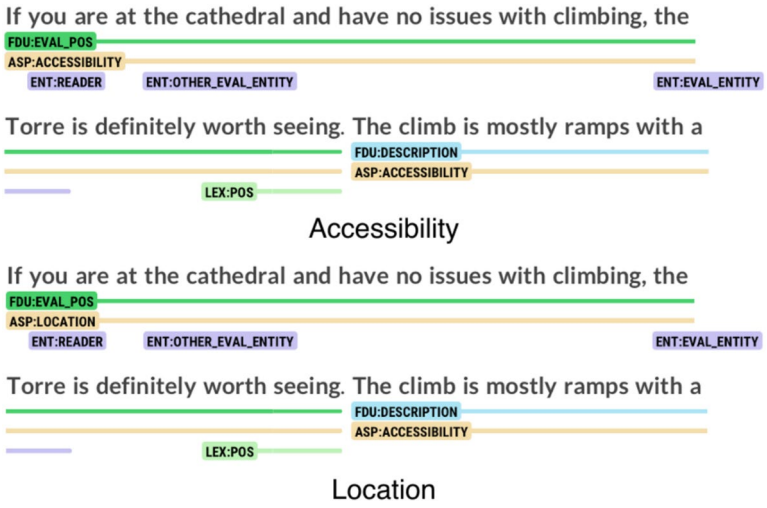

Fig. 11 Annotation of a segment as *accessibility* and *location*

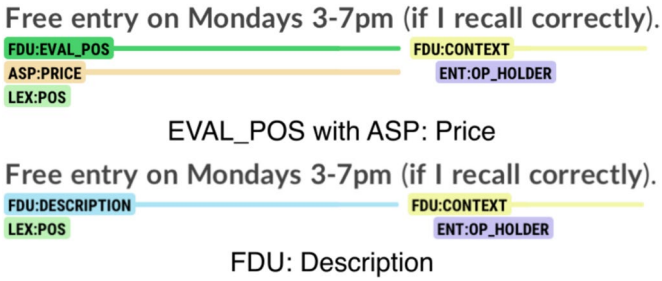

Fig. 12 Annotation of a segment as *eval_pos* or *description*

as *context*, interpreting it as a personal account of the reviewer's visit. In contrast, Annotators B and C labeled the same segment as a *positive evaluation*, most likely due to the presence of the adjective "amazing", which they interpreted as conveying an explicitly positive sentiment.

Aspects such as *accessibility* and *location* also posed challenges. Although *accessibility* was defined as "access to the monument," some annotators primarily associated it with facilities for disabled individuals, leading them to exclude instances that did not explicitly reference wheelchair users or similar accessibility concerns. Fig. 11 illustrates an example of such disagreement. While Annotators A and C classified the span as *accessibility*, Annotator B labeled it as location.

The annotation of text segments related to the price of the monument also presented challenges for annotators. In many cases, it was unclear whether the reviewer was evaluating the ticket price or merely providing factual information for the reader. Fig. 12 illustrates an example of such ambiguity. Annotator A classified the

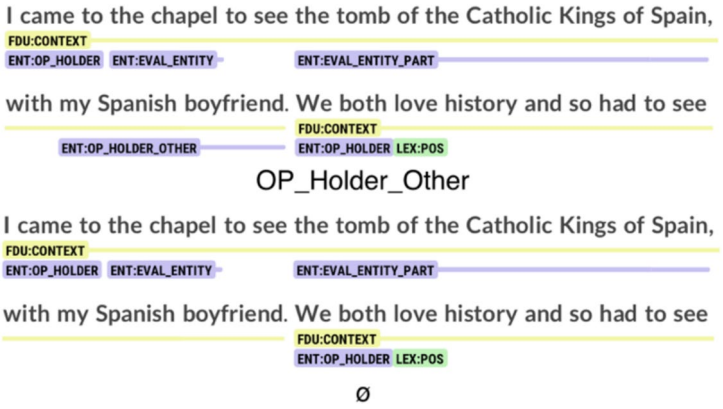

Fig. 13 Annotation of a segment as *op_holder_other* or nothing

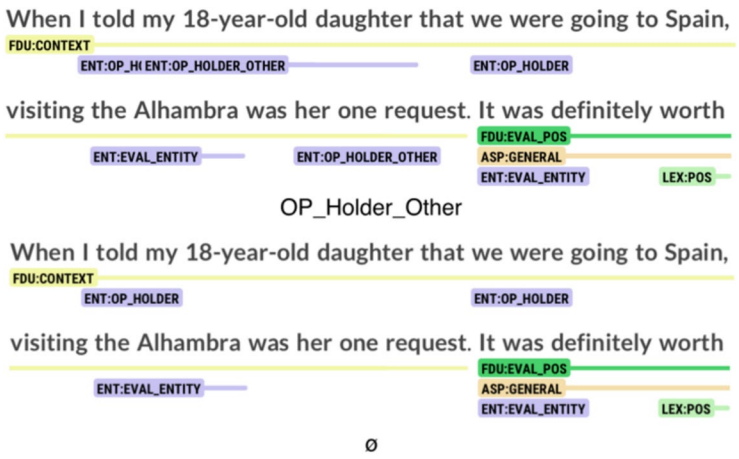

Fig. 14 Annotation of a segment as *op_holder_other* or nothing

text segment under the aspect price, likely influenced by the presence of the word "free". In contrast, Annotators B and C categorized the same segment as a descriptive FDU.

Although, as shown on Table 8, the layers concerning lexicon and entities presented less variation among annotators, they were not free from cases of disagreement. For the entities layer, a recurring issue involved the annotation of the opinion holder, particularly when using the plural "we". Fig. 13 illustrates an example of this challenge. Annotators A and B classified "my Spanish boyfriend" as *op_holder_other* and "we" as *op_holder*. However, Annotator C omitted the annotation of "my Spanish boyfriend", leading to a discrepancy in entity identification. A similar issue is observed in Fig. 14, where Annotators A and B tagged

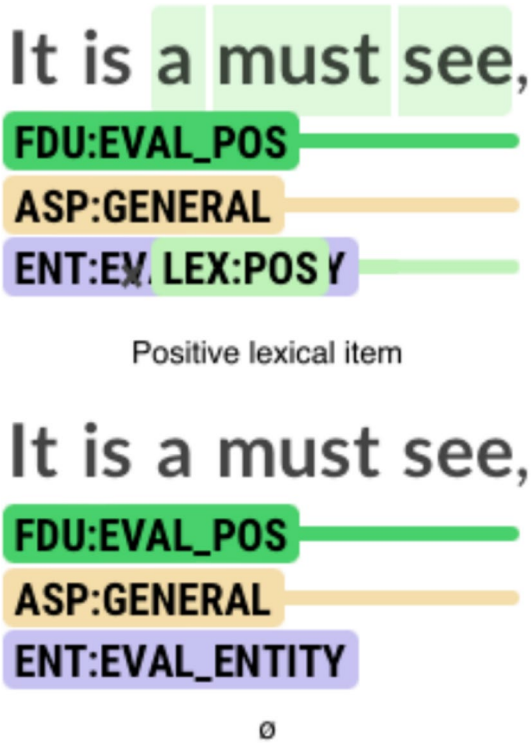

Fig. 15 Annotation of a segment as *lex_pos* or nothing

"my 18-year-old daughter" and "her" as *op_holder_other*, whereas Annotator C left them unannotated.

For the lexicon layer, one recurring issue among annotators involved the classification of text segments containing the expression "a must." As illustrated in Fig. 15, Annotators A and B classified the segment as a FDU of positive evaluation, recognizing "a must" as a positive lexical item, whereas Annotator C did not.

Conclusions

The quantitative results obtained from the annotation process represent one of the most valuable contributions of this study. Unlike most existing research, which often relies on comparatively smaller datasets and focuses primarily on qualitative analysis, our study follows a rigorous annotation procedure with a larger dataset, allowing for more robust empirical insights.

The final inter-annotator agreement metrics indicate substantial agreement among annotators across all four annotation layers, which suggests that the annotation schema was well understood and that annotators consistently applied the established criteria, despite the complexity of the schema and the density of the

annotation layers, and underlines the important role that proper guidelines applications and consensus sessions play in corpus annotation.

The results for the FDU layer indicate that positive evaluation is the most prevalent category, suggesting that users tend to highlight the positive aspects of a monument and express them openly. Additionally, users appear to favor sharing factual information or personal details about their visit rather than making negative statements. This pattern aligns with the findings for the lexicon layer, where positive lexical items accounted for 84.43% of the corpus, compared to just 15.57% for negative lexical items. These results suggest that when writing reviews, users generally emphasize the positive aspects of cultural attractions, employing words with positive semantic orientation while generally avoiding direct negative statements.

The most relevant aspect evaluated by reviewers, other than the general one, is "artistic value" (15.81%), "info" (8.18%), "price" (7.8%), "location" (6.78%), and "views" (5.97%), showing that these are the features that visitants tend to pay most attention to.

The qualitative analysis reveals several areas where the annotation schema can be refined to reduce ambiguity and enhance its overall effectiveness. Labels should be free of misinterpretations and ambiguity to ensure consistency among annotators. For instance, within the FDU layer, annotators frequently struggled to distinguish between the *context* and *eval_pos* tags when positive lexical items were included in statements. Similarly, in the aspects layer, the categories *accessibility* and *location* led to disagreement, as some annotators interpreted the former as applying exclusively to people with disabilities. The *price* aspect also proved to be a point of contention, as it was sometimes unclear whether the review author was evaluating the entry price or merely stating it as factual information.

To ensure the reliability of annotation schemas, it is crucial to establish clearer boundaries and precise definitions for labels, thereby minimizing ambiguity and reducing instances of annotator disagreement. These insights will be instrumental in refining the schema, ultimately contributing to more consistent and accurate analyses of user-generated content.

Finally, given the results achieved in this research, its main objective—to design, implement, and validate an annotation schema for ABSA with rich discursive features—may be said to have been achieved. While the annotation schema may undergo further refinement in future research, the findings provide (1) valuable insights into how evaluation is linguistically expressed in consumer reviews and (2) a foundation for fine-tuning a pre-trained Transformer model for the automatic identification of these discourse segments.

Acknowledgments This work was funded by the Spanish Ministry of Science and Innovation [grant number PID2020-115310RB-I00].

Author's Contribution All authors whose names appear on the submission made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; drafted the work or revised it critically for important intellectual content; approved the version to be published; and agree to be accountable for all aspects of the work in

ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding Funding for open access publishing: Universidad de Málaga/CBUA.

Data Availability The dataset used in this article is freely available and can be found at https://osf.io/rqfw4/.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Ethical Approval This is an observational study and it does not involve humans and/or animals.

References

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the international conference on recent advances in natural language processing*, pp. 1–7.

Benamara, F., Taboada, M., & Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics, 43*(1), 201–264. https://doi.org/10.1162/COLI_a_00278

Bhatia, V. K. (1983). Simplification v. Easification—the case of legal texts1. *Applied Linguistics, 4*(1), 42–54. https://doi.org/10.1093/applin/4.1.42

Biber, D., Connor, U., spsampsps Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins.

Cambria, E., Das, D., Bandyopadhyay, S., spsampsps Feraco, A. (2017). Affective computing and sentiment analysis. In E. Cambria, D. Das, S. Bandyopadhyay, spsampsps A. Feraco (Eds.), *A Practical Guide to Sentiment Analysis* (pp. 1–10). Springer International Publishing. https://doi.org/10.1007/978-3-319-55394-8_1

Cenni, I. (2024). Sharing travel experiences on TripAdvisor: A genre analysis of negative hotel reviews written in French, Spanish and Italian. *Journal of Pragmatics, 221*, 76–88. https://doi.org/10.1016/j.pragma.2023.12.015

De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., & Hoste, V. (2017). Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In *Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 136–142. https://doi.org/10.18653/v1/W17-5218

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. (2021). Identifying and describing functional discourse units in the BNC spoken 2014. *Text & Talk, 41*(5–6), 715–737. https://doi.org/10.1515/text-2020-0053

Ein-Dor, L., Shnayderman, I., Spector, A., Dankin, L., Aharonov, R., & Slonim, N. (2022). *Fortunately, Discourse markers can enhance language models for sentiment analysis* (arXiv:2201.02026). arXiv. https://doi.org/10.48550/arXiv.2201.02026

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Wiley.
Gaikwad, G., & Joshi, D. J. (2016). Multiclass mood classification on Twitter using lexicon dictionary and machine learning algorithms. In *2016 International conference on inventive computation technologies (ICICT)*, vol. 1, pp. 1–6. https://doi.org/10.1109/INVENTIVE.2016.7823247
Huber, P., & Carenini, G. (2020). From sentiment annotations to sentiment prediction through discourse augmentation. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 185–197). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.16
Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, pp. 7–36.
Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). *Twitter sentiment analysis: Lexicon method, machine learning method and their combination* (arXiv:1507.00955). arXiv. https://doi.org/10.48550/arXiv.1507.00955
Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. SAGE Publications.
Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data*. Springer.
Mann, W. C., spsampsps Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In G. Kempen (Ed.), In *Natural language generation: New results in artificial intelligence, psychology and linguistics* (pp. 85–95). Springer. https://doi.org/10.1007/978-94-009-3645-4_7
Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.
Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English.* Palgrave Macmillan.
Montani, I., Honnibal, M., Van Landeghem, S., Boyd, A., Peters, H., Samsonov, M., Geovedi, J., McCann, P. O., Regan, J., Orosz, G., Altinok, D., Kristiansen, S. L., Roman, Fiedler, L., Howard, G., Wannaphong Phatthiyaphaibun, Tamura, Y., Explosion Bot, Bozek, S., Henry, W. (2021). *Prodigy v1.11.4* (Version v3.1.0) [Computer software]. Explosion. https://doi.org/10.5281/ZENODO.1212303
Moreno-Ortiz, A. (2017). Lingmotif: Sentiment analysis for the digital humanities. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics*, (pp. 73–76). http://aclweb.org/anthology/E/E17/E17-3019
Moreno-Ortiz, A. (2024). The linguist's role in sentiment analysis: From knowledge provider to data annotator. In G. Garofalo & S. Maci (Eds.), *Investigating discourse and text. Corpus-assisted analytical perspectives* (pp. 25–54). Peter Lang.
Moreno-Ortiz, A., & Pérez-Hernandez, C. (2018). Lingmotif-lex: A wide-coverage, state-of-the-art lexicon for sentiment analysis. In *Proceedings of the 11th international conference on language resources and evaluation (LREC 2018)*, (pp. 2653–2659).
Moreno-Ortiz, A., Salles-Bernal, S., & Orrequia-Barea, A. (2019). Design and validation of annotation schemas for aspect-based sentiment analysis in the tourism sector. *Information Technology and Tourism, 21*(3), 535–557. https://doi.org/10.1007/s40558-019-00155-0
Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems, 108*, 92–101. https://doi.org/10.1016/j.knosys.2016.05.032
Pang, T. (2002). Textual analysis and contextual awareness building: A comparison of two approaches to teaching genre. In A. Johns (Ed.), *Genre in the classroom. Multiple perspectives* (pp. 145–161). Lawrence Erlbaum.
Panseeta, S., & Todd, R. W. (2014). A genre analysis of 5-star hotels' responses to negative reviews on tripadvisor. *rEFLections, 18*, 1–13. https://doi.org/10.61508/refl.v18i0.114196
Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryiğit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, (pp. 19–30). https://doi.org/10.18653/v1/S16-1002
Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, (pp. 486–495). https://doi.org/10.18653/v1/S15-2082
Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the 8th international*

workshop on semantic evaluation (SemEval 2014), (pp. 27–35). https://doi.org/10.3115/v1/S14-2004

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the 6th international conference on language resources and evaluation (LREC'08)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf

Prasad, R., Webber, B., & Lee, A. (2018). Discourse Annotation in the PDTB: The Next Generation. In H. Bunt (Ed.), *Proceedings of the 14th Joint ACL-ISO workshop on interoperable semantic annotation* (pp. 87–97). Association for Computational Linguistics. https://aclanthology.org/W18-4710

Rezvani Kalajahi, S. A., Neufeld, S., & Nadzimah Abdullah, A. (2017). The discourse connector list: A multi-genre cross-cultural corpus analysis. *Text & Talk*. https://doi.org/10.1515/text-2017-0006

Schouten, K., & Frasincar, F. (2016). COMMIT at SemEval-2016 Task 5: Sentiment analysis with rhetorical structure theory. In *Proceedings of SemEval-2016*, (pp. 356–360).

Shaukat, K., Hameed, I. A., Luo, S., Javed, I., Iqbal, F., Faisal, A., Masood, R., Usman, A., Shaukat, U., Hassan, R., Younas, A., Ali, S., & Adeem, G. (2020). Domain specific lexicon generation through sentiment analysis. *International Journal of Emerging Technologies in Learning (iJET), 15*(9), 190–204.

Swales, J. (1981). *Aspects of article introductions*. University of Aston.

Swales, J. (1990). *Genre analysis: English for academic and research settings*. Cambridge University Press.

Swales, J. M. (2004). Research genres: Explorations and applications. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139524827

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, (vol. 2, pp. 325–347). https://doi.org/10.1146/annurev-linguistics-011415-040518

Taboada, M., Voll, K., & Brooke, J. (2008). Extracting sentiment as a function of discourse structure and topicality. *Technical Report TR 2008-20*.

Thompson, G., & Hunston, S. (2000). Evaluation: An Introduction. In Thompson, Geoffrey & S. Hunston (Eds.), *Evaluation in text. Authorial stance and the construction of discourse* (pp. 1–27). Oxford University Press.

Upton, T. A., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes, 20*(4), 313–329. https://doi.org/10.1016/S0889-4906(00)00022-3

Van Dijk, T. A. (1977). *Text and Context: Explorations in the semantics and pragmatics of discourse*. Longman.

Vásquez, C. (2011). Complaints online: The case of Tripadvisor. *Journal of Pragmatics, 43*(6), 1707–1717. https://doi.org/10.1016/j.pragma.2010.11.007

Voll, K., spsampsps Taboada, M. (2007). Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. In M. A. Orgun spsampsps J. Thornton (Eds.), *AI 2007: Advances in artificial intelligence* (pp. 337–346). Springer. https://doi.org/10.1007/978-3-540-76928-6_35

Wang, B., & Liu, M. (2015). *Deep learning for aspect-based sentiment analysis.* Stanford University Report.

Zhang, Y., & Vásquez, C. (2014). Hotels′ responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context & Media, 6*, 54–64. https://doi.org/10.1016/j.dcm.2014.08.004