A protocol for the annotation of evaluative stance and metaphor across four discourse genres

Laura Hidalgo-Downing,[1] Paula Pérez-Sobrino,[2]
Laura Filardo-Llamas,[3] Carmen Maíz-Arévalo,[4]
Begoña Núñez-Perucha,[4] Alfonso Sánchez-Moya[4]
& Julia T. Williams Camus[5]
[1] Universidad Autónoma de Madrid | [2] Universidad de La Rioja |
[3] Universidad de Valladolid | [4] Universidad Complutense de Madrid |
[5] Universidad de Cantabria

The present article contributes to research on evaluation by addressing two complementary objectives: first, we present a protocol for the identification and annotation of evaluation in English discourse and, second, we show the results of the implementation of the protocol in the annotation of evaluation in a sample of a corpus of four genres. We first describe the protocol by discussing the theoretical and methodological grounding of the annotation scheme, the criteria, the categories, the steps for the implementation of the protocol and an illustrative example of the application of the protocol to a short extract. We subsequently provide the preliminary results of a pilot study with the frequency of evaluative expressions across the four genres. Results show that while adjectives and non-metaphoric evaluative expressions are overall more frequent, there are differences regarding the preference for positive or negative value and regarding the frequency of function.

Keywords: evaluation, stance, evaluative metaphor, evaluation annotation, evaluation in genres

1. Introduction

While the topic of evaluation has received increased attention in recent years, especially since the publication of the influential works by Hunston and Thompson (2000); Martin and White (2005) and Thompson and Alba-Juez

(2014), most research has focused on the discussion of theoretical issues concerning the definition and types of evaluation and on the analysis of evaluation markers in various discourse types. However, some issues concerning the study of evaluation, such as the identification and annotation of evaluation in discourse and the comparison of evaluative markers across genres, which we address in the present article, are still emerging areas of inquiry.

Regarding the identification and annotation of evaluation, recent publications have focused on the identification and annotation of Attitude categories within the Appraisal system and on the discussion of how these categories may be refined (see, for example, Fuoli, 2018; Fuoli & Hommerberg, 2015; Hidalgo Tenorio & Benítez de Castro, 2020; Read & Carroll, 2012; Taboada & Carretero, 2010; Thompson, 2014). A different approach to the identification and annotation of evaluation has been pursued by Simaki et al. (2018, 2020), who propose an utterance-based approach to speaker stance in a corpus of blog posts on Brexit. In this study, ten broad categories relevant for the expression of stance (including agreement/disagreement, certainty and contrariety, among others) are identified.

Different corpora and research objectives may require different procedures and methods for the identification of evaluation. A crucial issue in the identification of evaluation is the size of the corpus that is annotated and whether the main research objective is quantitative or qualitative. In this sense, there are two main trends in the research on evaluation: on the one hand, there are studies on the lexical expression of evaluation in small samples with detailed discussion of the categories (see, for example, Bednarek, 2009) and, on the other, studies of phraseological or grammatical evaluative patterns in large corpora (see, for example, Hunston, 2011; Hunston & Su, 2019). In our project on stance, our objective was to identify evaluative categories that can be easily retrieved electronically, for the purpose of both quantitative and qualitative analysis of the use of evaluation in the four genres. However, time limitations and the fact that the annotation of the corpus was carried out manually, led us to establish a set of ad hoc criteria and restrictions on the concept of evaluation (see Section 2.5. below).

One of the issues that has not received sufficient attention so far is the identification and annotation of evaluative metaphoric expressions, with Fuoli et al. (2021) being an exception. Indeed, although metaphor is mentioned by Martin and White (2005) in their Appraisal framework, studies of evaluative metaphor have typically been carried out from the perspective of Conceptual Metaphor Theory (see, for example, Deignan, 2010) but not from the perspective of stance and the expression of evaluation as a resource for the expression of stance. In our study we are interested in exploring the relation between the expression of evaluation and metaphoricity. This decision is supported by previous research (see, for example, Fuoli et al., 2021), which shows a direct correlation between the presence of lexical units with potential for metaphorical expression and the projection of evaluation over a certain topic.

Regarding research on evaluation across genres, this remains an under-researched area. Here, Biber et al.'s (1999) corpus comparison of the expression of stance markers in conversation, fiction, news, and academic discourse remains a crucial referent. However, Biber et al.'s (1999) approach does not focus solely on evaluation markers, but on stance markers in more general terms; these include, in addition to evaluative or attitudinal markers, markers of epistemicity and other markers of stance. Results in this study show that stance markers are more frequent in conversation than in the written genres, but that stance is still frequent in general terms in all genres, including academic discourse. The most prototypical category for the expression of evaluation is the adjective (Biber et al., 1999, pp. 512–513), though the frequency of the type of stance marker seems to vary depending on the specific genres. Adverbials, for example, are more frequent in academic discourse, and adjectives followed by complement clause are more frequent in news and fiction. Additional significant contributions to the study of evaluation across genres are found in the field of phraseology, with Hunston (2007) being a representative example; in this study she compares the pattern *it v-link ADJ that* in two corpora, one from *New Scientist* and another from *The Sun* and *News of the World*. Results of this study show that while the evaluative markers in the *New Scientist* focus on importance and likelihood as semantic categories, evaluative markers in the news reports concern the expression of judgements regarding what is acceptable or desirable. These studies reveal not only that the frequency of stance markers in various genres is different, but also that the type of marker that is favoured in each genre varies too. It follows that the investigation of the occurrence of evaluation and stance markers across genres is necessary in order to gain insights into the preferences of specific genres.

Against this background, the present article contributes to the study of evaluation in discourse by pursuing two main objectives: first, we present an annotation protocol developed for the identification and annotation of evaluation markers in a 400,000-word corpus of four different genres (newspaper opinion discourse, political discourse, fora and scientific popularization discourse) and, second, we show the preliminary results of the application of the protocol to the analysis of evaluation in samples from the corpus. For this purpose, we first describe in detail the protocol, together with the steps in the annotation procedure and the categories. We then present the results of a pilot study across genres, which will enable us to identify genre-specific trends regarding preferences for specific parts of speech (adjective, noun, verb, adverb), value (positive or negative) and metaphoricity (non-metaphoric versus metaphoric).

The article is organized as follows: Section 2 situates the present study within previous research on stance and evaluation and provides an in-depth description of the protocol, Section 3 provides an example of the application of the protocol to a short extract and Section 4 provides an overview of preliminary studies and results in the four genres. The article closes with conclusions in Section 5.

2. A protocol for the identification and annotation of evaluation in discourse

This section describes four main aspects of the design and implementation of the protocol for the identification and annotation of evaluation: (1) the theoretical concept of evaluation within stance and its definition as adopted in our scheme, (2) the motivations for the design of the protocol, (3) the main criteria for the identification and annotation of evaluative stance in discourse, (4) the description of the categories of the protocol, (5) the stages in the development and implementation of the protocol and (6) the steps in the annotation procedure.

2.1 Evaluation and stance: Theoretical grounding and methodological issues

Our annotation scheme draws from various sources and theoretical traditions, which our approach brings together so as to develop a rich approach to the concept of evaluation in discourse (Du Bois, 2007; Englebretson, 2007; Hunston & Tompson, 2000; Martin & White, 2005; Tompson & Alba-Juez, 2014). The relation between evaluation and stance remains a complex one, since some authors use the terms almost as synonyms (Hunston & Thompson, 2000; Martin & White, 2005) while others consider evaluation as one of the resources for the expression of stance, distinct from the areas of epistemicity-evidentiality and affect (Englebretson, 2007, p. 17). In this study, we adopt the latter position and consider stance a broader overarching category (see Alba-Juez & Thompson, 2014) within which evaluation is one of the three main resources for the expression of stance. We adopt the definition of stance found in Du Bois (2007, p. 163), who defines it as "a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field." In this view, stance as an interactive social act is inherently evaluative (also see Hunston & Thompson, 2000; Thompson & Alba-Juez, 2014) and involves the performance of acts of positioning of speakers and (dis)alignment with the addressees. However, while Du Bois's concept of stance overlaps with the concept of evaluation and encompasses the expression of attitude, epistemicity and affect, in our study, we restrict the analysis of evaluative language to the expression of opinions and attitudes as a specific area of enquiry within stance. This approach is supported by ample research on the expression of evaluation (see, for example, Bednarek, 2009). Indeed, while some scholars have proposed frameworks which integrate the expression of evaluation, epistemicity and affect (see, for example, Biber et. al, 1999; Hunston & Thompson, 2000; Martin & White, 2005; Thompson & Alba-Juez, 2014), most research on stance focuses on one of those three dimensions. These three areas of inquiry have distinct objects of study: evaluation is concerned with the expression of attitudes, opinions and assessments of entities and events; epistemicity and evidentiality are concerned with degrees of commitment towards propositions and reliability of sources of knowledge; and finally, affect is concerned with the expression of emotions and feelings (see, for example, Bednarek, 2009; Biber et al., 1999; Engelbretson, 2007). The three areas of inquiry are not clear-cut categories and, in fact, two or more of these meanings may co-occur in specific linguistic expressions, giving rise to complex stance-taking acts (Thompson & Alba-Juez, 2014, p. 7). The preliminary definition of evaluative stance we propose in our study is the following:

> The social act of assessing social actors, events and propositions by assigning positive or negative values which are grounded in socio-cultural systems of beliefs and opinions, and which express a subjective positioning of the speaker and intersubjective alignment or disalignment with specific communities of speakers.

Following this definition, the linguist's task as a discourse analyst will be to identify the expression of evaluation as the set of lexical resources which express positive or negative assessment of entities, events and propositions, which are open for examination by using linguistic analysis and methods.

2.2 Motivations for the design of the protocol

Much in line with previous protocols designed to annotate evaluation (see Fuoli, 2018; Simaki et al., 2018), our annotation scheme is concerned with the expression of evaluation by means of lexical markers. Working with a large corpus and annotating it manually did not make it possible to adopt the Attitude system within Appraisal Theory (Martin & White, 2005) and required more general ad hoc categories of evaluation. The design of our protocol is motivated by three main needs: first, the need to develop an annotation scheme of evaluation which will guarantee the use of unified criteria within a group of annotators working on a large corpus and the replicability of the identification and annotation procedure.

The protocol was designed for the purpose of tagging a 400,000-word corpus containing four subcorpora of 100,000 words each: political discourse (speeches delivered by British Conservative and Labour politicians), opinion discourse (*The Guardian* and *The Times*), press popularization articles (*The Guardian* and *The Times*) and fora discourse (REDDIT). The size and heterogeneous nature of the corpus made it necessary to restrict and narrow down our approach to the concept of evaluation and to elaborate ad hoc categories in order to make the annotation possible within a reasonable period of time and taking into account that the annotation was done manually. Second, it is important to bear in mind that the annotation of this corpus was part of the work carried out within a research project on stance in discourse,[1] which had the aim of analysing stance variation across discourse genres from a critical discourse perspective. The critical perspective motivated the collection of two subsets of data within the corpora of opinion discourse and political discourse, so that these corpora would allow for potential studies comparing evaluation strategies used by Conservative and Labour politicians and newspapers (see Hidalgo-Downing & Pérez-Sobrino, 2023; Núñez-Perucha & Filardo-Llamas, 2023). The critical perspective also motivated the use of a specific category of evaluation inspired in critical discourse analysis (see Section 2.4. below). Third, as pointed out in the introduction, a further concern was the identification of metaphoricity as a separate category in the annotation system, so that studies on evaluative metaphor in the various genres could be pursued in the future. In brief, our annotation system is motivated by the need to establish categories for the identification and manual annotation of evaluative expressions in a large corpus of four different genres, with a view to pursuing studies which can focus on different lexical categories of evaluation including evaluative metaphor from a critical discourse perspective.

2.3 Criteria for identifying evaluative stance

Four main criteria for the identification and annotation of evaluative stance are adopted. The word or phrase identified as expressing evaluative stance must meet the following conditions (following the definition of stance in Du Bois, 2007):

1. It must assess entities, events or propositions yielding positive or negative value, and it must be concerned with the expression of attitudes and opinions. In case of doubt regarding what counts as an evaluative expression, annota-

[1] The research project is entitled "Stance and Inter/Subjectivity in Discourse" (PGC 2018-095798), and is funded by the Spanish Ministry of Science, Innovation and Universities.

tors are asked to use Martin and White (2005) as a reference and to check the definitions of words in the *MacMillan* and *Collins Cobuild* dictionaries.[2]

2. It must express the positioning of the speaker/writer towards a topic or event in the discourse.
3. It must express alignment or disalignment with other voices or communities of speakers and shared systems of beliefs evoked in the discourse.
4. It must belong to one of the grammatical categories established in the scheme: NP, AP, ADVP, VP.
5. It must express first-degree stance. This means that, initially, our annotation excludes reported or second-degree stance as expressed in sentences included between inverted commas (the dimension of Attribution in Appraisal Theory). However, the analysis of second degree of stance is carried out in the case study on scientific popularization discourse illustrated in Section 2.6., given the important role played by this kind of stance in this genre.

Let us expand a bit further on some of the criteria outlined above. With regard to criterion 1, the identification of the evaluative expression as expressing a positive or negative assessment, this is possibly the most challenging step, as has already been pointed out above. In order to assist annotators in the clarification of possible doubts regarding whether an expression is evaluative or not, the *MacMillan* and *Collins Cobuild* dictionaries have been adopted as reference dictionaries to determine in a first instance whether the expression may be potentially evaluative or not.

In order for the word/expression to be annotated as evaluative, annotators will check whether the definitions include an evaluative connotation (either positive or negative). If the definition contains a connotation and this supports the intuition that the expression is evaluative in the observed context, it is annotated as evaluative. If the definition does not contain a positive or negative connotation, but it is still felt that the expression is evaluative, it will be checked carefully in context. Let us consider two practical examples, the potentially evaluative word *works* in the expression *a deal that works* and the word *revolutionise* used in a scientific text. Below is the definition of *work* in lexical entry 12 in the *Collins Cobuild* dictionary:

12. VERB

[2] The websites of the two dictionaries are: for MacMillan dictionary: https://www.macmillandictionary.com/; for Collins Cobuild dictionary: https://www.collinsdictionary.com/. The MacMillan Dictionary granted free online access when this article was written, but it was discountinued in 2023. The dictionaries were accessed online during October and November 2021.

If an idea, system, or way of doing something works, it is successful, effective, or satisfactory.

This definition of *work* includes three evaluative terms (*successful, effective, satisfactory*) which make explicit the positive connotation of the word. Words with inscribed evaluation (that is, explicit evaluation, such as *good* and *bad*) are usually evaluative by default and are annotated, except when the core meaning of the word has been lost in a specific context. The word *revolutionized* was found in the following sentence, from the science corpus: "the findings, which will be revealed at a conference at the Institute of Education in London this week, could revolutionise the treatment of anorexia". The definition of this word does not include a connotation in the dictionary; however, it was considered that its use in the context of scientific advancement had a positive evaluation and was consequently annotated as evaluative and yielding a positive value.

For the application of criterion 1, annotators are reminded that, when annotating evaluative expressions, words expressing epistemic and deontic stance are not annotated, even if this kind of stance also expresses a positioning and may express relations of alignment/disalignment. This means that stance expressions of possibility, probability, obligation, necessity or volition, which belong to the semantic domain of epistemic modality and evidentiality and not to the semantic domain of evaluation, are not annotated.[3] Annotators are also reminded that words which belong to the semantic domain known as Graduation within the Appraisal system are not annotated either, since these expressions are concerned with establishing scales of force and focus of evaluative expressions and are thus modifiers of evaluative expressions (see Martin & White, 2005).

Regarding the second and third criteria, that the evaluative expression should express a positioning of the speaker/writer and alignment/disalignment with communities of speakers or voices, it has to be noted that the relation between evaluation and positioning and dis/alignment is a complex one. Although Du Bois (2007) claims that all evaluative expressions indicate positioning and manifest relations of dis/alignment with other voices, his examples focus on interactional stance, while evaluation in written texts may work in different ways. Indeed, as the analysis of the extract in Section 3 below reveals, while positioning and dis/alignment is explicitly manifested by some expressions, it is not so prominent or even not relevant for value in others. It is possible that positioning and, in particular,

[3] The motivation for this is that in our research project, epistemic and deontic stance was annotated by a different group of researchers. However, it may be argued that some markers of epistemic and effective modality, when these take the form of adjectives and nouns and not modal verbs (for example *clear, possible, sure*), can also have an evaluative function. However, this multilayered analysis of stance was beyond the aims of the project.

dis/alignment are manifested at the discourse level through a cumulative process of evaluation, while value needs to be assigned individually for each term. This is the reason why it is important for annotators to read the complete texts before performing the annotation.

Regarding the fourth criterion, the unit of analysis we adopt is the word, and, when necessary, the phrase, for example in some metaphoric expressions. This means that longer stretches such as chunks and sentences are not annotated. We are aware that this imposes a limitation on the identification of evaluation, which is often expressed cumulatively, prosodically and in implicit ways, but it is worth noting that the interpretation and analysis of the evaluative expression will always be carried out in context. This means that annotators are expected to analyze and discuss the broader evaluative context of individual evaluative expressions. Typical words expressing evaluation are words such as *good, bad, disaster, problem, unluckily, (to) shrink, (to) divide*, including compound words such as *civil war*, hyphenated words such as *self-delusion* and phrasal verbs such as *break down*. In some cases, phrases are annotated, especially when annotating evaluative metaphors such as *chronic illness* (referring to racism, in *the Guardian*) and *civil war* (referring to internal conflicts in the Tory party, in *The Guardian*). The criterion adopted is that there should be a unity of meaning in the expression which is different from the sum of the individual lexical items.

2.4 Categories

Our protocol for the identification and annotation of evaluation consists of four categories: part of speech, function, metaphoricity and value. Each of these is described below.

2.4.1 Part of speech

For doubts regarding part of speech identification, the annotation protocol included numerous examples of the grammatical categories which typically convey evaluative stance, inspired in Biber et al.'s (1999) stance categories and Hidalgo-Downing's (2016) overview of the relation between evaluation and grammar. The adjective was the most straightforward category for the identification of evaluation. Good examples of this category from our corpus are expressions with inscribed evaluation, such as *right, ridiculous, shameful, odd* (opinion), *better, optimistic, healthy, strong, true* (politics), *frustrated, hard, short-sighted, difficult, hypocritical* (fora), *significant, important, vital* (science). The protocol reminds annotators of specific types of adjectives that are potential markers of evaluation. These include adjectives denoting fairness (*fair, understandable*) because they imply agreement or support (alignment), and personal interpretations or reformulations of technical data in scientific discourse, such as a *majo*r study, *serious* health problems (science). An interesting case concerns adjectives such as *true* and *false*. The criterion adopted in our annotation is that examples such as *this is true* involve evaluation if there is a division of opinions about the statement (e.g. in the context of fake news) and consequently, a positioning is being expressed, and not a fact.

The identification of evaluative nouns, verbs and adverbs is more complex. Regarding nouns and verbs, the main difficulty has to do with the fact that nouns and verbs may perform two different functions, a function of categorization, according to which an entity or event is categorized in terms of a given class, and a function of evaluation, according to which an entity or event is evaluated by means of expressing an opinion about or attitude towards that entity (van Dijk, 1995, p. 29). Van Dijk illustrates this double function by drawing attention to the possible uses of the word *thief*, which may be used to categorize an entity according to a class as established by the law, and is thus understood as describing a fact, or may be used to express an opinion about an entity, in which case it does not describe a fact. Good examples of evaluative nouns in our corpus are expressions with inscribed evaluation, such as *prejudice* (fora), *disaster* (opinion), *problem, success* (opinion and politics) and metaphoric expressions such as *psychodrama, boardgame, precipice* (opinion). Similarly, good examples of evaluative verbs in our corpus include verbs with inscribed evaluation such as *winning* (politics), *lobbying* (opinion), and metaphoric expressions such as *plundered, embrace, build* (politics), *suffers* (opinion), *ripped off* (fora). With regard to evaluative adverbs, the least frequent category in our corpus, good examples include attitudinal sentential adverbs which take scope over the whole proposition, such as *happily, unfortunately* (opinion), manner adverbs which modify verbs, such as for example *doggedly* (politics), *vigorously, irresistibly, bitterly, desperately* (opinion) and adverbs which modify adjectives, such as for example *naturally* (in "most naturally gifted politicians", opinion).

2.4.2 Function

Our annotation scheme includes three functions, classifying (CLA), predicational (PRE) and attitude (ATT). The categories CLA and PRE are inspired in nomination and predicational strategies in critical discourse analysis (henceforward CDA) (van Leeuwen, 2008; Wodak & Meyer, 2015), while the category ATT is inspired in the category of attitudinal stance adverbs in Biber et al. (1999). The motivation for including these categories is to be able to group, first, the evaluative representation of entities and events (CLA); second, the evaluation of qualities and properties of these entities and events (PRE); and, third, the evaluation of whole propositions (ATT). The predicational strategy (PRE) is concerned with the expression of a quality or property of a social actor or event relevant to the topic of a text, based on socio-cultural values and beliefs. The prototypical part of speech that realizes the predicational strategy is the adjective or adjective phrase.

The classifying category is based on what different authors call categorization, classification or the function of referential/nomination strategies in CDA (van Leeuwen, 2008; Wodak & Meyer, 2015). An entity is classified according to a specific variable, typically, race, ethnicity, geographical or social origin or role, etc. Classification is typically expressed in nouns and NPs but is also applicable to verbs, which can be said to categorize events. Most evaluative metaphors will typically have a classifying function, but some may occur in the predicational strategy, as modifiers of other categories (see, for example, *glittering* career, opinion). Note, however, that not all classifying nouns are evaluative.

The attitude category is inspired in Biber et al.'s (1999) category of attitudinal adverbials and is concerned with the expression of an attitude towards a proposition, rather than towards an entity or event. The prototypical realization of attitude is the sentential adverb or adverb phrase, such as for example: *happily*, *unfortunately* (press). Notice, however, that this function can also be performed by adjectives, typically in *it + to be + AP* structures, including those which have undergone ellipsis. Examples of this structure which we codify as ATT are the following: "it's *worth* asking why", "*True*, America's pro-life movement has plenty of prominent women in it" and "But, *sad* to say, this time Labour is not one of them".

2.4.3 Metaphoricity

In order to determine whether an evaluative expression is metaphoric or not, two criteria are applied: first, the distinction between contextual and basic meaning postulated by the Metaphor Identification Procedure (MIP) (Praggeljaz, 2007, later revised by Steen et al., 2010 in MIPVU) is adopted. For this purpose, metaphoricity is checked by annotators in the *MacMillan* Dictionary. If the meaning of a word is listed as one of the basic entries of the dictionary, typically lexical entry 1, it is not annotated as metaphoric (there are, however, exceptions). Second, there must be a mapping from a source domain (typically, though not necessarily) with a positive or negative value onto a target domain. Two examples of terms that are not annotated as metaphoric are the words *key* and *vital* (science), as confirmed by the first definitions in the *MacMillan* dictionary. In cases such as these, it can also be argued that the potential mapping between a source and a target domain is no longer activated and they are what is known in the literature as dead metaphors.

As a general rule, highly conventionalized metaphors (again, this needs to be checked for degree of metaphoricity in dictionaries) are not annotated. For example, the word *gutted* in "I'm *gutted* that Northumbrian is now banned" (fora) is annotated as non-metaphoric after confirming its meaning coincides with lexical entry 1 in the *MacMillan* dictionary. Further examples of non-evaluative metaphors from our corpora include expressions such as: *quarters* of the Christian community, *entering* the toughest phase of negotiations. These expressions are highly entrenched or conventionalized metaphors and additionally we considered they lack a positive or negative value.

Let us consider the two examples in (1) and (2), which are similar because they belong to the same extended metaphor IDEOLOGY IS A JOURNEY/LANDSCAPE. That is, both make use of a mapping from a source domain that belongs in the JOURNEY/LANDSCAPE metaphor onto an abstract target domain. The context is that the writer is drawing a comparison between gun supporters and Islamic terrorists and their presence on the internet.

(1) Tommy Robinson, Alex Jones, Paul Joseph Watson have **travelled** the exact route pioneered by their Islamist forebears.

(2) **Exiled** now to the web boondocks ….

In (1), the expression *travelled*, which has neither positive nor negative connotations, might sound evaluative, because in this context it could be interpreted as a periphrasis of "becoming radicalized", which could be annotated as evaluative. But the word "travelling" in itself does not express the writer's positioning towards the topic or alignment/dealignment with communities of speakers. For this reason, the expression is not annotated as evaluative. In (2) the expression *exiled* shows a mapping from a source domain which carries a negative connotation (someone who has been forced to live in a foreign country because they cannot live in their own country, usually for political reasons) onto the target domain of being restricted to places in the web that are difficult to find. It can also be interpreted that, by using this negative term, the writer is expressing his own positioning to the topic. The writer could have said "X has now moved to Y place", which would then be like the "travelling" example.

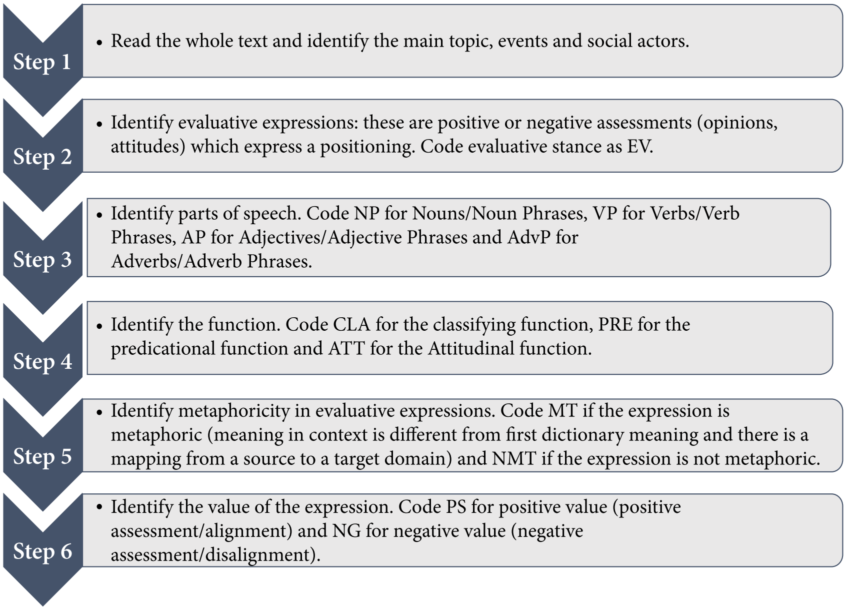
The prototypical part of speech for evaluative metaphor is the noun, followed by the verb, the adjective and the adverb. Good examples of evaluative metaphors are, first, metaphors which are not highly conventional and involve processes of de-agentivization and de-humanization of social actors, and second, metaphors which highlight the positive or negative effects of events. For example, Brexit as a topic is represented in opinion articles by means of expressions such as *cliff-edge*, *precipice*, *gang-plank* and *divorce*. These metaphors are used to highlight potential negative effects of Brexit. Other evaluative metaphors involve word category shift, such as "*poisoning* the well of political discourse" (opinion, from noun to verb) and highly incongruous collocations, such as "the Brexit *rainbow*" (opinion), which are also instances of creative uses of metaphor. Other examples of evaluative metaphors are verbs, such as *erode* the natural capital (opinion) and adjectives, such as their "*glittering* careers" (opinion) and "a *strong* economy needs a *healthy* environment" (politics).

2.4.4 Value

Three criteria are applied for determining value: first, that the expression acquires negative or positive value in context; second, that the expression carries a positive or negative connotation which expresses the speaker/writer's positioning in context; and third, that the speaker/writer's positioning evokes relations of dis/alignment with other voices or communities of speakers. While criteria 1 and 2 are applied in all cases (that is, the evaluative expression must have a positive or negative connotation in context which expresses the speaker/writer's positioning), it is not always possible to apply the third criterion; that is, it is not always possible to determine the speaker/writer's dis/alignment with communities of speakers by considering only individual evaluative items. Indeed, relations of dis/alignment more typically emerge prosodically in a cumulative way. Annotators are reminded that positive and negative evaluation is performed by a speaker/writer of the text towards an object/entity/event mentioned or evoked in the discourse. It is not the evaluation the analyst assigns to that object/entity/event, it is not the value that a potential addressee may assign, and it is not necessarily the value that object/entity/event may have in society at this moment in time. In case of doubt, annotators are reminded to check the definitions in the dictionaries mentioned above, especially regarding the specification of explicit positive or negative connotations of words. An additional possibility is to check concordances in corpora to try to determine whether a specific expression has a tendency to occur with positive or negative connotations. Annotators are also reminded that some terms have inscribed (explicit) evaluation such as *good, bad, right, wrong, true, lie, biased, problem*. However, not all positive or negative terms are also evaluative as pointed out in Section 2.6. below.

2.5 Annotation procedure

Regarding the steps in the annotation procedure, annotators are instructed to follow the steps outlined below:



Step 1

- Read the whole text and identify the main topic, events and social actors.

Step 2

- Identify evaluative expressions: these are positive or negative assessments (opinions, attitudes) which express a positioning. Code evaluative stance as EV.

Step 3

- Identify parts of speech. Code NP for Nouns/Noun Phrases, VP for Verbs/Verb Phrases, AP for Adjectives/Adjective Phrases and AdvP for Adverbs/Adverb Phrases.

Step 4

- Identify the function. Code CLA for the classifying function, PRE for the predicational function and ATT for the Attitudinal function.

Step 5

- Identify metaphoricity in evaluative expressions. Code MT if the expression is metaphoric (meaning in context is different from first dictionary meaning and there is a mapping from a source to a target domain) and NMT if the expression is not metaphoric.

Step 6

- Identify the value of the expression. Code PS for positive value (positive assessment/alignment) and NG for negative value (negative assessment/disalignment).

Figure 1. Steps in the annotation procedure

2.6 Stages in the development and implementation of the protocol

We outlined a preliminary annotation procedure based on theoretical definitions of evaluation and its main categories as outlined in the preceding sections. The preliminary model was tested on a 4862-word sample (featuring a balanced distribution of all genres under scrutiny) that was randomly extracted from the 400,000-word corpus, in 4 subsequent rounds of individual annotation of the samples following the protocol.[4] For the sake of practicality, only three researchers took part in the first and second rounds, and the whole team of seven researchers

[4] An inter-rater reliability test was conducted among the three researchers involved in all four rounds of analysis on two tasks: the identification of evaluative units and the classification of such units according to the categories of analysis. The results of the inter-rater reliability tests show high degree of agreement regarding the identification of evaluative units of analysis (ranging from F-score = 0.78 between researcher 2 and 3 to F-score = 0.86 between researchers 1 and 2) and a consistent increase in the Fleiss Kappa scores for the value category (positive vs. negative evaluation, from $k = 0.84$ in the first round to $k = 1$ in the fourth round), and, to a lesser extent, for metaphoricity (without any variation between the first and fourth round, $k = 0.79$). For a more detailed information about the inter-rater reliability process, see Hidalgo-Downing and Pérez-Sobrino (2022). The sample of texts used for the four rounds of annotations, statistical scripts, and annotated datasets by researcher and by round are available in a publicly accesible repository: https://osf.io/c2x6m/.

were involved in the third and fourth rounds in order to make sure that the annotation criteria were clear and easy to implement by researchers not involved in the definition of the protocol. After each round, the researchers met to discuss their annotations, debate diverging annotations and clarify instructions of the protocol where needed.

The main issues that were debated during the four rounds, which led to refinements in the annotation protocol, were the following: the first and main source of debate involved the identification of evaluative expressions, that is, the application of step 2, which, however, was closely linked to the identification of value (step 6). For a clarification of what counts as an evaluative expression, it was crucial to adopt the criterion that the lexical item must express positive or negative value. This meant that descriptive terms which are often part of the jargon of a specific genre (often in politics and science) were discarded. Examples include terms such as (Brexit) *deal, serve* (politics), *harm, suffer, decline* (talking of bees in science popularization) and also expressions related to contrariness or difference (*contrary, different, distinct*) which were not used to express positive or negative connotation. A related issue that was discussed was the concept of value (step 6). Here it was essential to establish a distinction between negative and positive terms and evaluative terms (see Bednarek & Caple, 2013). Positive and negative terms such as *war* and *peace* are not necessarily evaluative, in fact they are often not evaluative. The evaluative potential of the word or expression can only be determined in context. In order to determine whether the potentially evaluative terms had a positive or negative value in context, it was crucial to apply the criteria mentioned above, that is, not only that the term had a positive or negative connotation, but also that it expressed the speaker/writer's positioning, and, whenever possible, dis/alignment with communities of speakers.

A second issue was the distinction between markers of evaluation and markers which form part of the Graduation system in the Appraisal model (see Martin & White, 2005). This was particularly significant for the identification of evaluative adjectives and adverbs (step 3).

A third issue of debate concerned the identification of evaluative metaphors (step 4). Difficulties in identifying evaluative metaphors had to do mostly with the two following issues: (1) to what extent highly conventionalized metaphors such as *far* in *far*-right (press) can be considered as evaluative, and (2) the difficulties in determining how many expressions are evaluative metaphors in cases such as extended metaphor over a stretch of discourse. With regard to issue (1), the criterion adopted was that the definitions of the expressions in the dictionary should contain explicit indications of the connotations (positive or negative) of the words. If we consider the expression *far* in *far-right*, it is first annotated as metaphoric, since there is a discrepancy between the meaning of *far* in context (lexical entry 4b "used for talking about how extreme someone's actions are or how great an effect they have") and the more basic meaning ("used for talking about distance"). Additionally, this definition and the examples provided in the dictionary suggest that the value is negative. Indeed, the expression *far-right* is not used by members of the far-right, but only by other political positions that consider this ideology too extreme. If connotations do not form part of the definition of the word or these connotations are not clear in the examined context, these terms are discarded. With regard to issue (2), we may find that metaphoric expressions cluster in a part of the text, creating an extended metaphor. On these occasions, an initial metaphor opens an evaluative frame which is then enriched with subsequent terms. An example is found in a headline from the science corpus: "The critically *ill* NHS needs a big cash *injection*". Both *ill* and *injection* are annotated as evaluative metaphors, since they are analyzed as forming part of an extended metaphor related to illness. This example shows that the value of a metaphoric expression might arise in context, and not just because there is a mapping of a source domain with positive or negative value.

The final issue of debate had to do with the identification of value (step 6) when the evaluative term was under the scope of negation, in which case the value underwent reversal. For example, in the sentence "He was not *glamorous*", the value of "not *glamorous*" is negative and is annotated as negative, not as positive.

3. Implementing the protocol: An example

In this section an illustration is provided of how the protocol has been implemented in the identification and annotation of evaluation in an extract from an article from *the Guardian* on Brexit. This extract is from the opinion discourse sample described in Section 4. below. The paragraph is reproduced below together with the codes assigned by the annotators in the annotation process. Only representative evaluative expressions from this extract are explained below.

> Brexit is a *disaster* <EV, NP, CLA, NMT, NG> in the making, a *reactionary* <EV, AP, PRE, NMT, NG> project that will *damage* <EV, VP, PRE, MT, NG> our economy, *narrow* <EV, VP, CLA, MT, NG> the horizons of future generations, *shrink* <EV, VP, CLA, MT, NG> Britain's influence and *curb* <EV, VP, CLA, NMT, NG> our ability to cooperate on *grave* <EV, AP, PRE, NMT, NG> and urgent questions that go beyond national boundaries, that it will make life *harder* <EV, AP, PRE, NMT, NG>, *not easier* <EV, AP, PRE, NMT, NG>, for those who most *desperately* <EV, AdvP, PRE, NMT, NG> need change (*The Guardian*, EOG-1).

The whole text is first read to identify the main topic and social actors. The article, from *The Guardian*, is entitled "Remain voters are left with no choice but to ignore Labour next week" and discusses the position of Remain voters in the wake of the European elections in 2019. The position of the author is highly critical towards Brexit. Potentially evaluative words are coded EV, to distinguish it from epistemic and deontic stance in our corpus. The first word that is identified as evaluative is *disaster*, which is annotated as a NP, performing the function CLA, given that it is the head of a NP. To check whether the noun is metaphoric and to confirm the negative value the *MacMillan* dictionary is consulted. Here lexical entry 1 of this word is:

Disaster: something very bad that happens and causes a lot of damage or kills a lot of people

This definition coincides with the meaning of the word in context in the analyzed paragraph, so the word is annotated as not metaphoric (NMT). The negative value is confirmed by its connotations and is annotated as NG.

The second word is *reactionary*, which is identified as an adjective and is coded as AP and PRE, since it premodifies a noun and expresses a quality regarding that noun. The meaning is checked in the *MacMillan* dictionary to confirm it is not metaphoric and is annotated as NMT, and the value in context is determined as negative and is coded as NG. Notice that there are three further evaluative adjectives in the text, *grave*, *harder* and *not easier*, and that *easier* in this context is annotated as expressing negative evaluation (NG), because it falls under the scope of negation.

The next four words are verbs. The first three, *damage*, *narrow* and *shrink* are coded as metaphoric, while *curb* is coded as non-metaphoric. These four words are first annotated as verbs and as performing the classifying (CLA) function because of their position in the clause and because they describe events related to the topic of the text. Each of the verbs is checked for metaphoricity and value. *Damage* is not listed as a verb in the *MacMillan* dictionary, so it is checked in the *Collins Cobuild* dictionary, where the first lexical entry defines this verb as follows: "To damage an object means to break it, spoil it physically, or stop it from working properly". Because the meaning of *damage* in the context of this extract is different from the meaning of the lexical entry, since it does not describe the physical damage of an object, it is annotated as metaphoric (MT). Because of the connotations it is also annotated as having negative value. With regard to the verb *narrow*, the *MacMillan* dictionary does not provide entries for *narrow* as a verb, while the *Collins Cobuild* dictionary does not provide a lexical entry for *narrow* as synonymous of *restrict*, the meaning in this context. For these reasons, *narrow* is annotated as metaphoric. The co-text (*narrow the horizons*) provides a clue for the annotation of the expression as negative (NG). The verb *shrink* is defined in lexical entry 1 in the *MacMillan* dictionary as follows: "to become smaller in size". The meaning of this verb in this context does not coincide with this definition but rather means "decrease", so it is annotated as metaphoric. The contextual connotations of the decrease of Britain's influence point at a negative value and the expressions is annotated as negative. The verb *curb* is annotated as not metaphoric (NMT) because the definition of this word in lexical entry 1 in the *MacMillan* dictionary corresponds to the meaning of the verb in this context: "to control or limit something that is harmful". The definition and the use of the word in context also confirm the negative connotation of the word, which is annotated as having negative value.

Finally, the word *desperately* provides an example of an evaluative adverb. It is classified as an adverb because of its position and role in the clause as a modifier of the verb (Adverb of manner). As such, it is annotated as performing a predicational function (PRE). The definition in the *MacMillan* dictionary confirms the expression is not metaphoric (NMT) and has negative value (NG).

To finish the analysis of this extract, we can reflect briefly on the way in which the cumulative choices of evaluative words discussed above contribute to the expression of a specific stance towards the main topic, Brexit. All the words in this extract carry negative value; as such, they are used to express a very critical stance and positioning of the writer towards the discussed event, Brexit (described as a *disaster*). At the same time, some of these choices create explicit relations of disalignment with the persons responsible for or supporting Brexit (a project called *reactionary* and described as a *disaster*) and highlight the negative effects it may have on Britain's future (described as *narrowing* its horizons, *shrinking* Britain's influence and making life *harder*, for example).

4. Variation of evaluation across genres: A pilot study

In this section we present the preliminary results of the implementation of the protocol in samples from the corpora of the four discursive genres. The main objectives of this pilot study are the following: (1) to present preliminary findings in each of the genres and to determine differences and similarities in the frequency of the evaluation categories across the four samples; (2) to discuss preferences for specific choices regarding evaluation in each of the genres.

4.1 Data

The data for this preliminary study consists of three samples (newspaper opinion discourse, scientific popularization discourse, political discourse) and the complete corpus of fora discourse. In their complete form, each of the genres, collected between 2016 and 2018, consists of around 100,000 words: political discourse (speeches delivered by British Conservative and Labour politicians), opinion discourse (*The Guardian* and *The Times*), press popularization articles (*The Guardian* and *The Times*) and fora discourse on social issues (REDDIT). The data used for the present pilot study are as follows:

- A sample of 20 articles of opinion newspaper discourse from the British newspapers *The Times* (9,893 words) and *The Guardian* (10,575 words) on Brexit (20,468 words).
- A sample of political speeches delivered by Conservative (15,537 words) and Labour (14,546 words) politicians at party conferences between 2016–2019 (30,083 words).
- A sample of 15 science popularization articles from *The Guardian* (11,859 words) and *The Times* (7,088 words) (18,947 words).
- The complete corpus of fora discourse, collected from REDDITT, on social issues (103,749 words).

4.2 Method

For the identification and annotation of the evaluation categories in the samples, the protocol described in Section 2.5. above was implemented by pairs of researchers working on each genre, following the procedure described in Section 2.6. above. The categories in the annotation scheme (part of speech, function, metaphoricity, value) were quantified in order to obtain frequency graphs which show preliminary tentative results.

4.3 Results and discussion

With regard to the frequency of evaluative words in the four genres, higher frequencies are displayed in opinion (*N*=937, 4.57%) and political discourse (*N*=1360, 4.52%), followed by fora (*N*=2067, 1.99%), and science popularization (*N*=206, 1,08). These results confirm preliminary expectations regarding the potentially more evaluative nature of some discursive genres such as opinion and politics and the less evaluative nature of scientific newspaper discourse. Further research is needed to explore why the discourse of fora displays a relatively low frequency of evaluation.

The distribution of the categories of evaluation in the four genres is shown in Figure 2.

Regarding part of speech, as expected, the adjective is overall the most frequent evaluative category, with a higher frequency in fora (55.06%) and science (53.40%), followed by politics (50.96%), and press (opinion) (43.65%), while the adverb is the less frequent category, with a similar distribution across genres (between 3.05% in fora and 3.40% in science). The noun is the most variable category, with the highest frequency in opinion press (35.01%), followed by science (32.52%), politics (27.50%) and fora (23.90%). Further research is necessary to explore why there is such variation of evaluative nouns and whether this variation is found in larger corpora. Verbs show a similar frequency across genres (around 18%), except science (10.68%), where it is much lower. In terms of function, results show a predominance of the predicational strategy across genres, being most frequent in fora (67.05%), followed by science (55.83%), politics (54.93%) and opinion press (45.89%), while the classifying strategy is more frequent in opinion press (53.26%), followed by politics (44.78%), science (43.20%) and fora (31.79%). Regarding metaphoricity, evaluation is expressed most frequently in a non-metaphoric way, especially in fora. Metaphoric evaluation is more frequent in science (28,16), opinion press (28.07%) and politics (26.25%), than in fora (13.06%). Finally, results show that value is highly dependent on genre: positive value is more frequent in science (57.28%) and politics (52.28%) while negative value is more frequent in opinion press (75.35%) and fora (59.55%).

4.4 Examples from the different genres: Preliminary results from case studies

This section presents an overview of the preliminary results of case studies on evaluation in the four genres mentioned above, and in which the annotation scheme described above has been applied.

4.4.1 Newspaper opinion discourse

The annotation protocol was applied to the identification and annotation of evaluative stance in a case study whose aim was to analyze how choices in evaluative stance contribute to the construal of the news value of negativity in the two samples (20,468 words) described above, of articles from the British newspapers *The Times* and *The Guardian* on Brexit (see Hidalgo-Downing & Pérez-Sobrino, 2023). The comparison of the two newspapers, a conservative and a progressive one, allowed for a critical discourse perspective of the implications of choices in evaluative stance. The objectives of this study were the following: (1) to compare the frequency of lexical units expressing evaluation in each newspaper (2) to dis-

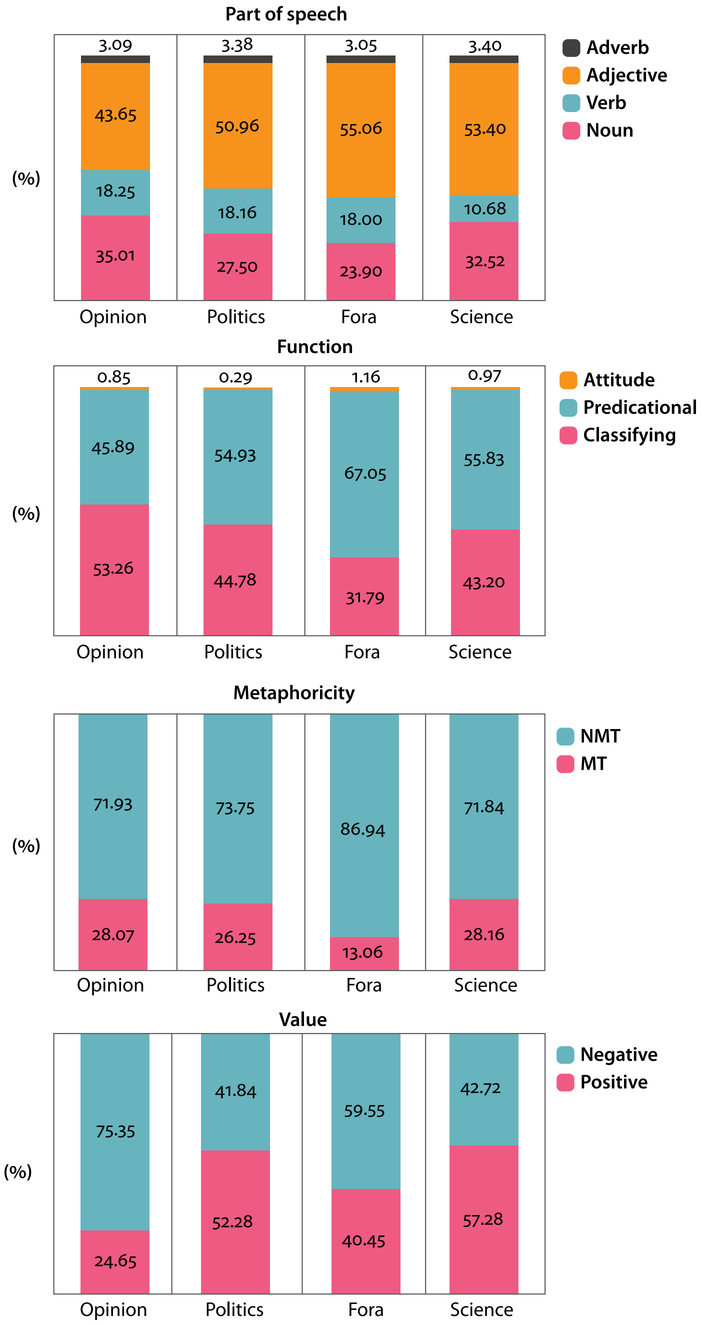


Figure 2. Proportion of evaluative categories in the four genres

cuss how negative evaluative stance towards Brexit is expressed in the whole corpus and in each of the newspapers, both of which are pro-remain. Figure 3 below shows the results regarding frequency and distribution of evaluative stance markers in the two newspapers.

Results show that regarding part of speech and function, although the adjective is the most frequent part of speech, the classifying function (which contains nouns and verbs) is more frequent than the predicational function, in particular in *The Guardian*. Differences between the newspapers can be observed regarding the frequency of evaluative metaphors and negative value, which are more frequent in *The Guardian* than in *The Times*. On examination, these choices revealed that *The Guardian* expressed an overall harsher criticism of Brexit as an event and of the social actors involved and a much more negative construal of the whole process. In *The Times* the expression of negative value focused on the possibility of a no deal Brexit rather than on the process itself. While *The Guardian* used more dramatic metaphors such as *cliff edge* or *precipice* to refer to Brexit, *The Times* used the less dramatic *divorce* metaphor. Some of these metaphors with negative value in *The Guardian* were also creative, confirming results found in the literature regarding the tendency for creative metaphors to express negative rather than positive value (Fuoli et al., 2021). Some of the representative negative evaluative expressions used to assess and construe Brexit and politicians involved in the process are the following: *bad, pointless, terrible, hard, stupid, irrational* (non-metaphoric adjectives), *problem, conundrum, bigotry, dishonesty, deceit, failure, disaster* (non-metaphoric nouns), *devour, purge, civil war, precipice, cliff-edge, divorce* (metaphoric nouns and verbs).

4.4.2 Political discourse

This case study examined the expression of evaluative stance in the sample of British Conservative and Labour speeches described above (30,083 words). The speeches were delivered between 2016 and 2019, a period of Conservative government (see Núñez-Perucha & Filardo-Llamas, 2023). Specifically, it aimed to analyse the frequency of parts of speech, function, metaphoricity and value of evaluative markers as explained in the annotation protocol above.

The quantitative and qualitative analysis of the two samples of speeches has revealed more similarities than differences. Figure 4 displays the overall results in terms of frequency and distribution of evaluative markers.

Results show that, by categories, the most frequent one is predicational evaluation, primarily realized by means of adjectives, closely followed by classifying expressions, particularly in the Conservative sample. Regarding figurativity and value, non-metaphoric markers and positive evaluation were the most recurrent types in both sets. Metaphoric evaluation was found to be mostly encoded by

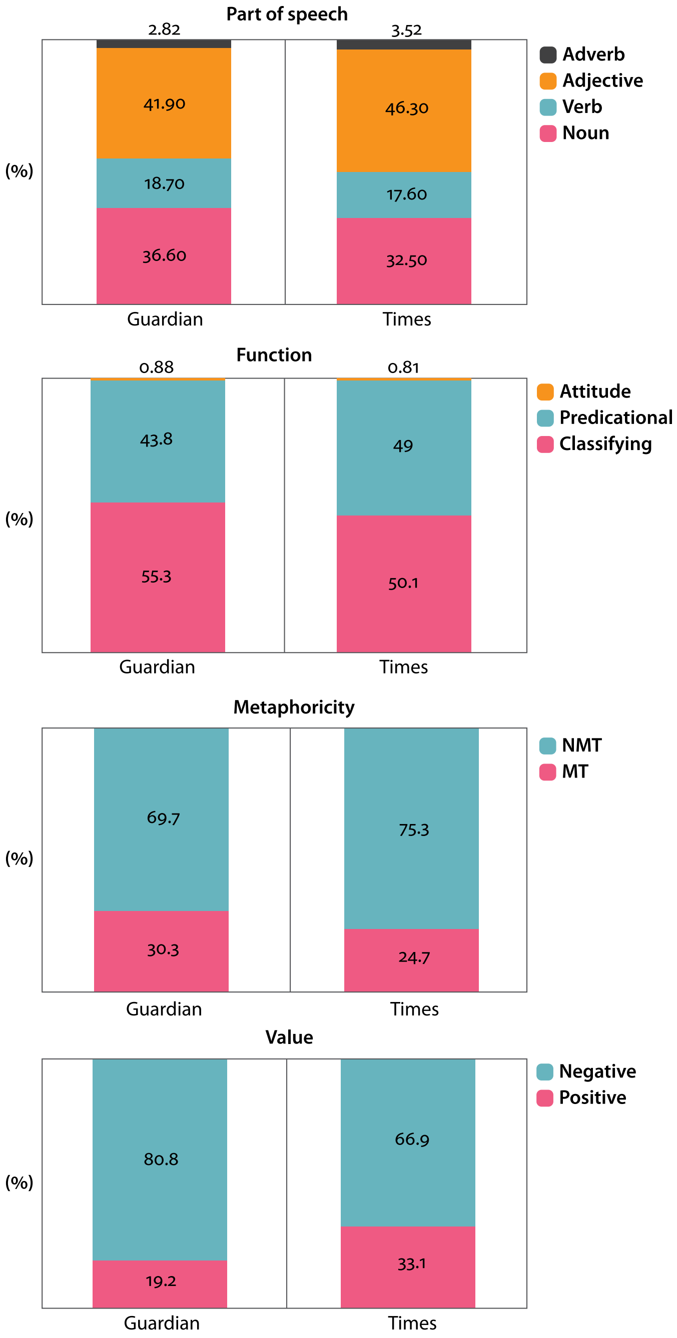


Figure 3. Distribution of evaluation categories in the two newspapers of the press opinion corpus

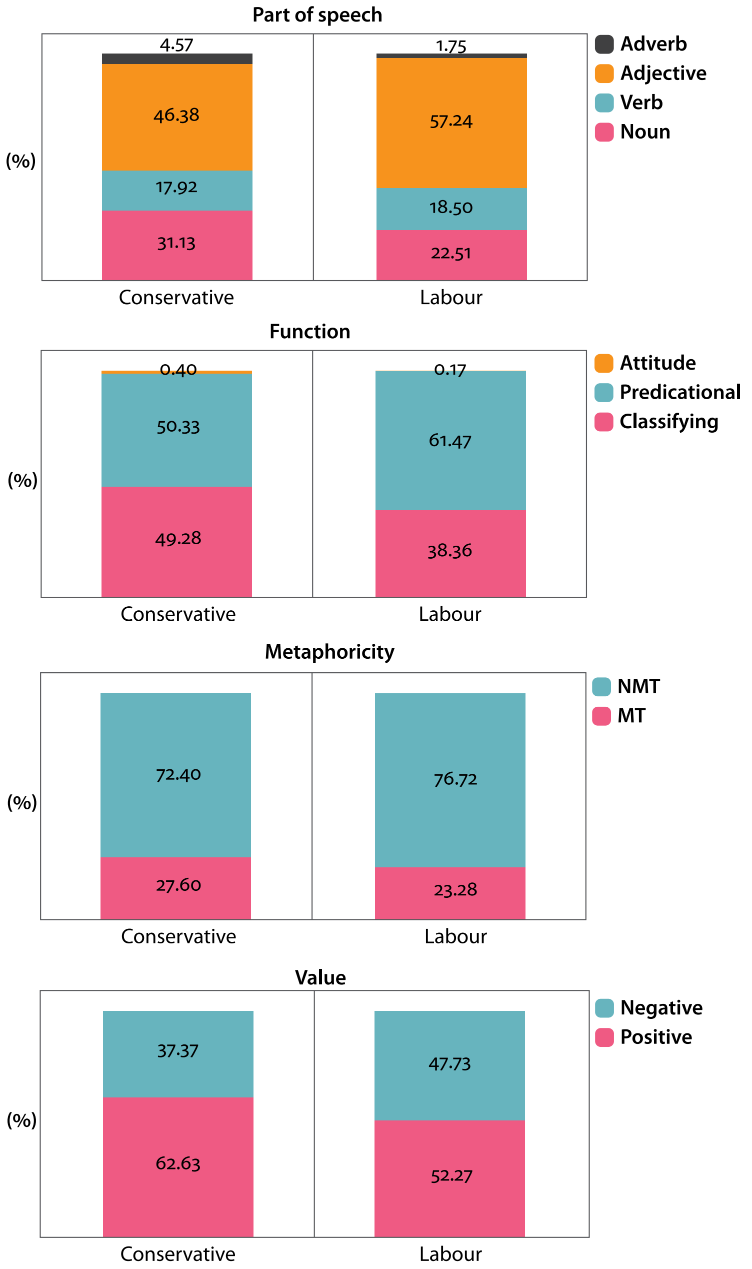


Figure 4. Distribution of evaluation categories in the British Labour and Conservative sets of the political discourse corpus

nouns and verbs with a classifying function as well as typically associated with conventional metaphors invoking the source domains of BUILDING, WAR and MOVEMENT, the former being more frequently used by the Labour party to represent themselves as agents of the transformation that Britain needs. Creative uses of metaphor were also present in both sets, although to a lesser extent, and were found to convey negative evaluation of the opponent. Overall, the findings revealed that, regardless of the speaker's ideological orientation, the type and value of stance markers are inscribed within a discursive strategy of positive self-presentation and negative other representation, thus contributing to legitimizing the party's ideological views and actions. Some of the representative evaluative expressions used to represent positively the self and the party (and negatively the other) are the following: *threat, opportunity, scandal* (non-metaphoric nouns), *essential, good, great* (non-metaphoric adjectives), *fight, (re)build, enemies, step, strength* (metaphoric nouns and verbs). Specific uses related to political discourse include non-metaphoric adjectives in the comparative (*greater, better, worse*) or the superlative form (*greatest, biggest)*, and the use of party names and ideologies (*Conservative, extremist*) with an evaluative function.

4.4.3 Press popularization articles

This study analyzed evaluative stance and metaphor in a comparable sample corpus of 30 popularization articles drawn from a larger corpus of 164 texts on scientific advances (Williams Camus, 2023). The sample corpus consisting of 15 texts with 11,859 words (*The Guardian*) and 7,088 words (*The Times*) was analyzed to identify the evaluative expressions, which, following the general protocol described above, were tagged for part of speech, function, metaphoricity and value. Although the other genres examined in the research project restricted the analysis to first-degree stance, the significant presence of reported speech in science popularizations made it necessary to include second-degree stance in the analysis.

The aims were to determine (1) the most frequent categories employed; (2) the preference for metaphoric or non-metaphoric language and (3) the source of evaluation – journalists (first-degree) or scientists and other stakeholders (second-degree). Figure 5 shows the results in *The Guardian* and *The Times*.

The results showed that the predicational function, generally realized by adjectives, was the most frequent category, followed by nouns with a classifying role, with only a marginal presence of attitudinals. As shown in Figure 5, there was a preference for non-metaphoric over metaphoric terms. For value, positive evaluation outnumbered negative stance. In general, there were only slight differences between the newspapers; however, there was a fairly large difference between positive and negative value in *The Guardian* whereas in *The Times* the difference

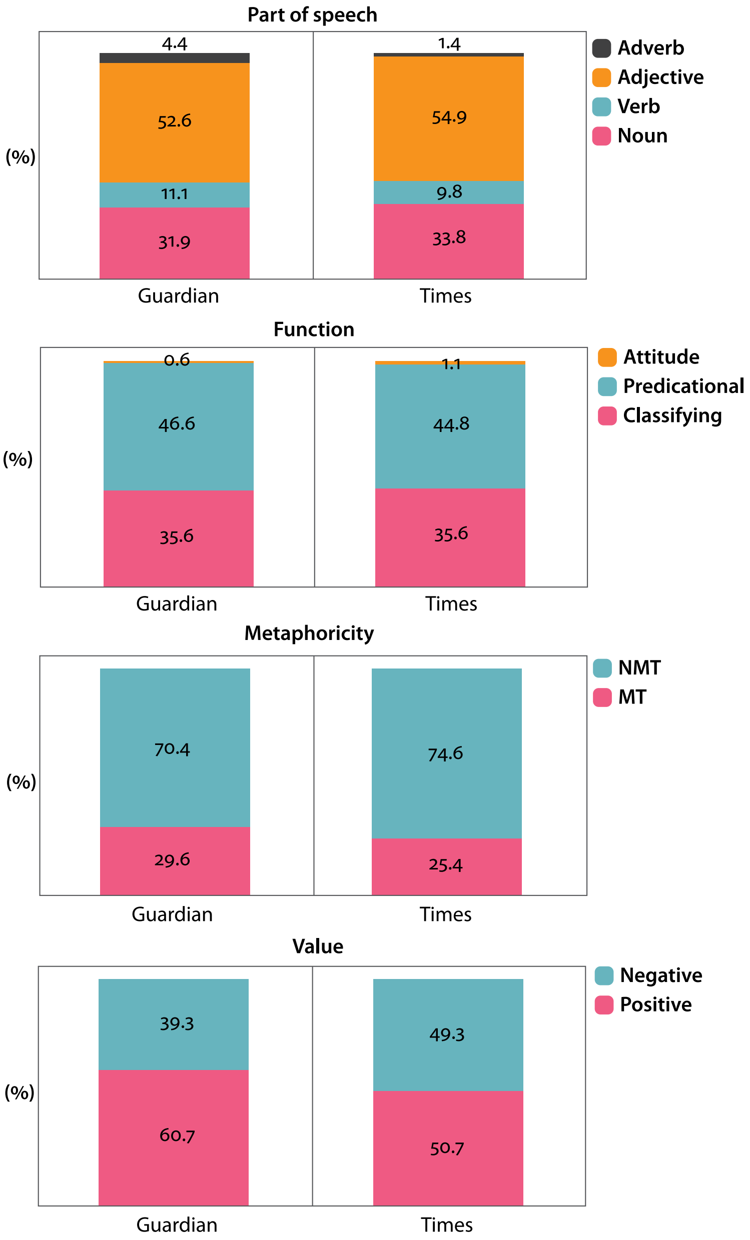


Figure 5. Distribution of evaluation categories in the two newspapers of the science popularizations corpus

was small. With regard to source of evaluation, both newspapers showed a preference for second-degree over first-degree stance: 90 vs. 47 per 10,000 words in *The Guardian* and 78 vs. 45 in *The Times*. Thus, overall, evaluation in science popularizations tends to be realized through non-metaphoric adjectival phrases of positive value in a predicational role and drawn from external (i.e. second-degree) sources. Some prototypical examples of positive adjective phrases include *important, vital, key, crucial* or *significant*, which are used to highlight the degree of importance of a scientific achievement. Other positive adjective phrases were also used to express a number of science-related issues, such as to indicate an emotional response (*enthusiastic, exciting, striking*), to signal the originality and/or applicability of the research (*radical, first; usable, valuable*) or to indicate ethical suitability (*clear, permissible, not unacceptable*).

4.4.4 Fora

This case study analyzed evaluative stance in a corpus of 28 online fora threads about a wide range of topics (Maíz-Arévalo & Sánchez-Moya, 2023). The total number of words for this *ad hoc* corpus was 103,749. Forum posts were gathered from co.uk domains to provide uniformity in the variety of English under scrutiny. To comply with ethical standards in online research, forum threads were publicly accessible and not password protected. The main objective was to ascertain the frequency and distribution of evaluative markers in digital fora discourse following the annotation protocol described above. See Figure 1 above for the distribution for each of the categories in this corpus.

According to our analysis of the dataset at hand, evaluative stance in online forums is lexically instantiated via predicational strategies (67.05% of the total), which grammatically corresponded to adjectival phrases (55.06%). Classifying strategies were found to amount to 31.79% and attitudinal markers were the least frequent ones (1.16%). Metaphoric evaluation in this sample was identified in 13.06% of the total cases, and value is more frequently negative (59.55%). Overall, then, evaluation in online fora is expressed in non-metaphorical, adjectival phrases with negative value. In fact, this study contributes to similar research in these lines (Jaki et al., 2019; Lorenzo-Dus & Nouri, 2020; Prażmo, 2020) on the type of evaluative value found in this set of forum threads. The predominance of negative value might also be explained by some of the affordances (namely anonymity) users seek when engaged in communicative exchanges in online forums (Prażmo, 2020). Some representative examples of evaluative expressions with negative value are the adjectives *buffoonish*, referring to an interpretation in a film, and *histrioni*c (referring to the film adaptation of a musical), negative metaphors such as *a stab at the police*, "Moulin Rouge was a major *turkey*" and positive metaphors such as "welcome to the *mind feast* of Climate-debate.com" and "UK looks like a renters *paradise*." Some negative evaluative expressions were highly creative, such as for example the noun *re-pubic-lick-uns*.

5. Conclusions

This article has presented an annotation protocol for the identification and annotation of evaluative stance in discourse together with examples of its implementation in four different discourse genres. The annotation scheme has been explained in detail, together with the criteria which were adopted for the definition of evaluative stance and each of the subcategories, and the steps followed in the annotation process. Given that identifying and annotating evaluative stance is a challenging endeavour, our article contributes to advancing research in the methodological difficulties involved in annotating evaluation, and to clarifying some controversial issues involved in the annotation of evaluation (for example, the notion of evaluative metaphor and the concept of value). We are aware that the annotation scheme has limitations for the potential identification of evaluative language resulting from the restricted criteria which have been imposed for practical reasons, especially concerning the unit of analysis (word or phrase), the restriction of value to positive and negative and the restriction to first-degree stance in three of the genres.

However, we believe that the protocol is adequate for the annotation of evaluative stance in different discourse varieties, as the case studies have shown. The results of the four studies reveal that there are similarities across the four genres, with non-metaphoric evaluation and the Adjective as a Part of Speech as the most frequent resources for the expression of evaluation. These results confirm similar results obtained in previous research and our expectations regarding our data. However, there were also interesting results which point at genre specific choices. First, regarding the category of function, opinion showed a preference for the classifying function, while the other three genres showed a preference for the predicational strategy. Second, the preference for positive or negative value seems to be motivated both by topic and by genre. Thus, positive value is favoured by political discourse to express the positive presentation of self against the backdrop of the negative representation of opponents, and by science popularization to highlight advances in scientific discoveries and theories. By contrast, negative value is favoured by opinion discourse on Brexit and by fora discourse on social issues to express negative criticisms of the various issues addressed. Third, evaluative metaphors are more frequent in science, opinion and politics, but less frequent in fora. These results point to the need to investigate in greater depth the motivation for different choices and preferences in each of the evaluation categories in different discourse genres.

Funding

The present study was carried out within the research project entitled "Stance and Inter/Subjectivity in Discourse" (PGC2018-095798), funded by the Spanish Ministry of Science, Innovation and Universities. Funding has also been granted from other research projects funded by the State Agency of Research – Spanish Ministry of Science, Innovation and Universities (FFI2017-82730-P, PID2020-118349GB-I00 and PID2021-123302NB-I00) and the Government of Aragón (LMP143_21).

References

Alba-Juez, L. & G. Thompson. (2014). The many faces and phases of evaluation. In G. Thompson & L. Alba-Juez (Eds.) *Evaluation in context* (pp. 3–24). John Benjamins.

Bednarek, M. (2009). Language patterns and ATTITUDE. *Functions of Language*, *16*(2), 165–192.

Bednarek, M., & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society*, *25*(2), 135–158.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman Pearson.

Deignan, A. (2010). The evaluative properties of metaphors. In G. Low, Z. Todd, A. Deignan, & L. Cameron (Eds.) *Researching and applying metaphor in the real world* (pp. 357–374). John Benjamins.

Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.) *Stancetaking in discourse* (pp. 139–182). John Benjamins.

Englebretson, R. (Ed.). (2007). *Stancetaking in discourse*. John Benjamins.

Fuoli, M. (2018). A stepwise method for annotating Appraisal. *Functions of Language*, *25*(2), 229–258.

Fuoli, M., & Hommerberg, C. (2015). Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora*, *10*(3), 315–349.

Fuoli, M., Littlemore, J., & Turner, S. (2021). Sunken ships and screaming banshees: Metaphor and evaluation in film reviews. *English Language and Linguistics*, *26*(1), 1–29.

Hidalgo-Downing, L. (2016). Grammar and evaluation. In C. Chapelle (Ed.) *The Applied Linguistics Encyclopedia*. Online publication. [Accessed 2 October 2021].

Hidalgo-Downing, L., & Pérez-Sobrino, P. (2022). Developing an annotation protocol for evaluative stance and metaphor in discourse: Theoretical and methodological considerations. *Text and Talk*. Advanced online publication.

Hidalgo-Downing, L., & Pérez-Sobrino, P. (2023). 'Pushing Britain off the precipice': A CDA approach to negative evaluative stance in opinion articles on Brexit. In J. Marín-Arrese, L. Hidalgo-Downing, & J.R. Zamorano Mansilla (Eds.) *Stance, inter/subjectivity and identity in discourse* (pp. 201–226). Peter Lang.

Hidalgo Tenorio, E., & Benítez Castro, M.Á. (2020). The language of evaluation in the narratives by the Magdalene laundries survivors: The discourse of female victimhood. *Applied Linguistics*, *42*(2), 315–341.

Hunston, S. (2007). Using a corpus to investigate stance quantitatively and qualitatively. In R. Englebretson (Ed.) *Stancetaking in discourse* (pp. 27–48). John Benjamins.

Hunston, S. (2011). Corpus Approaches to Evaluation. *Phraseology and Evaluative Language*. Routledge.

Hunston, S., & Su, H. (2019). Patterns, constructions, and local grammar: A case study of evaluation. *Applied Linguistics*, *40*(4), 567–593.

Hunston, S., & Thompson, G. (2000). *Evaluation in text*. Oxford University Press.

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, *7*(2), 240–268.

Lorenzo-Dus, N., & Nouri, L. (2020). The discourse of the US alt-right online–a case study of the Traditionalist Worker Party blog. *Critical Discourse Studies*, *18*(4), 1–19.

Maíz-Arévalo, C., & Sánchez-Moya, A. (2023). "Histrionic, appalling, a major turkey": The expression of evaluative stance in the discourse of online forums. In J. Marín Arrese, L. Hidalgo-Downing, & J.R. Zamorano Mansilla (Eds.) *Stance, inter/subjectivity and identity in discourse* (pp. 249–269). Peter Lang.

Martin, J.R., & White, P.R.R. (2005). *The language of evaluation. Appraisal in English*. Macmillan.

Núñez-Perucha, B., & Filardo-Llamas, L. (2023). From "roaring lion" to "chlorinated chicken": Evaluative stance and ideological positioning in a corpus of British political discourse. In J. Marín-Arrese, L. Hidalgo-Downing, & J.R. Zamorano Mansilla (Eds.) *Stance, inter/subjectivity and identity in discourse* (pp. 227–248). Peter Lang.

Pragglejazz Group (2007). A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, *22*(1), 1–39.

Prażmo, E. (2020). Foids are worse than animals. A cognitive linguistics analysis of dehumanizing metaphors in online discourse. *Topics in Linguistics*, *21*(2), 16–27.

Read, J., & Carroll, J. (2012). Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, *46*(3), 421–447.

Simaki, V., Paradis, C., & Kerren, A. (2018). Evaluating stance-annotated sentences from the Brexit Blog Corpus: A quantitative linguistic analysis. *ICAME journal*, *42*(1), 133–166.

Simaki, V., Paradis, C., Skeppstedt, M., Sahlgren, M., Kucher, K., & Kerren, A. (2020). Annotating speaker stance in discourse: The Brexit Blog Corpus. *Corpus Linguistics and Linguistic Theory*, *16*(2), 215–248.

Steen, G., Dorst, Aletta G., Herrmann, J.B., Kaal, A.A., & Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics*, *21*(4), 765–796.

Taboada, M., & Carretero, M. (2010). Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in English and Spanish. *Linguistics and the Human Sciences*, *6*(1–3), 275–295.

Thompson, G. (2014). Affect and emotion, target-value mismatches, and Russian dolls: Refining the appraisal model. In L. Alba-Juez, & G. Thompson (Eds.) *Evaluation in context* (pp. 47–66). John Benjamins.

Van Dijk, T.A. (1995). Power and the news media. In D.L. Paletz (Ed.) *Political communication and action* (pp. 9–36). Hampton Press.

van Leeuwen, T. (2008). *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford University Press.

Williams Camus, J.T. (2023). Evaluative stance in science popularisations in the English press. In J. Marín-Arrese, L. Hidalgo-Downing, & J.R. Zamorano Mansilla (eds.) *Stance, inter/subjectivity and identity in discourse* (pp. 271–293). Peter Lang.

Wodak, R. & M. Meyer (Eds.) (2015). *Methods of Critical Discourse Studies*. Sage.

Resumen

Este artículo contribuye a la investigación sobre la evaluación al abordar dos objetivos complementarios: en primer lugar, presentamos un protocolo para la identificación y anotación de la evaluación en el discurso en lengua inglesa; y, en segundo lugar, mostramos los resultados obtenidos al implementar dicho protocolo en una muestra de un corpus de cuatro géneros discursivos. Comenzamos con la descripción del protocolo y la fundamentación teórica y metodológica del mismo, así como los criterios, los pasos para la implementación del protocolo y un ejemplo ilustrativo de su aplicación a un texto breve. A continuación, presentamos los resultados preliminares de un estudio piloto sobre la frecuencia de las expresiones evaluativas en cuatro géneros discursivos. Los resultados muestran que los adjetivos y las expresiones evaluativas no metafóricas son más frecuentes, pero hay diferencias en los diferentes géneros en la frecuencia de la evaluación positiva o negativa y la frecuencia de las funciones.

Palabras clave: evaluación, posicionamiento, metáfora evaluativa, anotación de la evaluación, evaluación en géneros discursivos