Annotating Evaluative Language: Challenges and Solutions in Applying Appraisal Theory

Jiamei Zeng
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR
jiamezeng3-c@my.cityu.edu.hk

Min Dong
School of Foreign Languages
Beihang University
PR China
mdong@buaa.edu.cn

Alex Chengyu Fang
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR
acfang@cityu.edu.hk

Abstract

This article describes a corpus-based experiment to identify the challenges and solutions in the annotation of evaluative language according to the scheme defined in Appraisal Theory (Martin and White, 2005). Originating from systemic functional linguistics, Appraisal Theory provides a robust framework for the analysis of linguistic expressions of evaluation, stance, and interpersonal relationships. Despite its theoretical richness, the practical application of Appraisal Theory in text annotation presents significant challenges, chiefly due to the intricacies of identifying and classifying evaluative expressions within its sub-system of Attitude, which comprises Affect, Judgement, and Appreciation. This study examines these challenges through the annotation of a corpus of editorials related to the Russian-Ukraine conflict and aims to offer practical solutions to enhance the transparency and consistency of the annotation. By refining the annotation process and addressing the subjective nature in the identification and classification of evaluative language, this work represents some timely effort in the annotation of pragmatic knowledge in language resources.

Keywords: Appraisal Theory, Attitude, evaluative language, pragmatic annotation

1. Introduction

Appraisal Theory (Martin and White, 2005) describes a taxonomy of semantic resources that allow for the expression of emotions, judgements, and valuations as well as the means to enhance and engage with these evaluations (Martin 2000, p.145). It has attracted an increasing academic interest evidenced by a growing volume of publications in the Web of Science (Figure 1), indicating the urgent need for the pragmatic analysis of evaluative language.

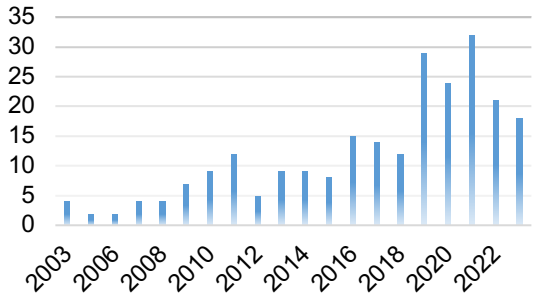

Figure 1: Annual count of academic publications on Appraisal Theory from 2003 to 2023

Considered as a systematic, detailed and elaborate framework for the analysis of evaluative language (Bednarek, 2006, p. 32), Appraisal Theory has demonstrated a great expanding relevance across various fields including, among many others, the examination of academic discourse (e.g. Swain, 2010; Hood, 2010; Geng and Wharton, 2016), political language (e.g. Mayo and Taboada 2017), news narratives (e.g. Bednarek and Caple, 2010; Huan, 2016), business discourse (e.g. Pounds, 2011; Fuoli and Hommerberg, 2015), wine tasting sheets (Breit, 2014), movie reviews (Taboada et al., 2014), and public statements (Meadows and Sayer, 2013).

However, as a sophisticated analytical framework involving semantic and pragmatic interpretations, the theory is not without its challenges, particularly when applied to the annotation of large corpora of natural texts. A major challenge lies in the dual tasks of annotation practices: identifying textual elements of appraisal and classifying them according to the theory's component categories of Attitude, Engagement, and Graduation and their respective sub-categories (Fuoli, 2018). This complexity is compounded by the inherent subjectivity and variability of linguistic expressions.

Fuoli (2018) suggests a step-wise method as a general solution. Our work to be reported next aims to provide more detailed solutions by targeting the Attitude category and addressing specific problems and issues, thereby exploring the issue of operationality through clear, operable strategies. In particular, we constructed a corpus of editorials from news media, performed the annotation of this material according to the Attitude system, and reviewed the various problematic issues before the formulation of solutions. We aim to offer additional insight about the aspects of applying a theoretically rich but operationally challenging framework through practical annotation of a sound level of transparency and consistency. We also hope that efforts such as ours will help to harness the full

ISA-20 Proceedings @LREC-COLING-2024, pages 144–151

potential of Appraisal Theory for the analysis and understanding of evaluative language.

2. Methodological Issues

This section provides a comprehensive outline of the methodological framework applied in the appraisal annotation of editorial content. We will first explain the rationale behind the selection of editorials as the primary material and introduce the composition of the annotator team. Following this, an in-depth examination of the chosen annotation framework, the tool utilized, and the procedural steps undertaken will be presented. These elements collectively form the foundation of our systematic approach.

2.1 Corpus Data

A corpus was constructed comprising editorials, selected for their inherent nature of presenting opinions, making them an ideal subject for this study. Four diverse newspapers were selected as the primary sources of data, including China Daily (CD), New York Times (NYT), South China Morning Post (SCMP), and The Guardian (TG). Thirty editorials were selected from each newspaper, all of which were published between January 2022 and May 2023 and centred on the Russian-Ukraine conflict, amounting to a total of 120 articles. The corpus of editorials is summarized in Table 1. This time frame and subject matter were set up to capture a wide range of evaluative perspectives during a period of significant geopolitical tension.

	CD	NYT	SCMP	TG	Total
Text	30	30	30	30	120
Token	14,073	15,170	20,975	18,551	68,769
Type	2,982	3,368	4,255	4,035	8,678

Table 1: Summary of the corpus of editorials

2.2 Annotation Framework

Appraisal System is defined as the linguistic mechanisms through which authors or speakers express their positive or negative assessments regarding the subjects, events, and situations discussed in their texts (Martin and White, 2005, p. 2). It is divided into three primary systems: Attitude, Engagement, and Graduation, each with its own sub-systems or categories. Our annotation experiment focused on the Attitude system, which comprises Affect (emotional responses), Judgement (evaluations of human behaviour and character), and Appreciation (assessments of objects, texts, events, and processes). Each dimension features a polarity aspect, allowing classifications as either positive or negative.

Affect is the core sub-system of Attitude and is subdivided into four categories: Dis/inclination, Un/happiness, In/security, and Dis/satisfaction. Judgement is divided into two sub-systems including social esteem and social sanction. Social esteem relates to the evaluation of someone's abilities (Capacity), their adherence to norms (Normality), and their persistence or determination (Tenacity). Social sanction focuses on truthfulness (Veracity) and appropriateness or morality (Propriety). Appreciation evaluates reactions to, compositions of, and valuations of objects or phenomena.

2.3 Annotators and Annotation Tool

The annotation of the corpus was performed by six MA students in linguistics, divided into three annotation groups with two annotators each. UAM Corpus Tool (O'Donnell, 2008) was chosen as the annotation tool for the experiment. It has a user-friendly interface and provides modules for statistical analysis of the annotated data. This feature was useful for the presentation and interpretation of our annotation results.

2.4 Annotation Process

The annotation process involved the initial training of the annotators to ensure a sound level of consistency measured in terms of inter-annotator agreement before the full-scale annotation of the corpus was rolled out. The process involved the following specific steps:

Step 1: Each group were first of all required to familiarize themselves with Martin and White (2005) in general and Attitude in particular during the first stage.

Step 2: A tutorial session was given to all the annotators, key concepts summarized and major principles outlined. An annotation guide was drawn up.

Step 3: A first trial annotation was performed simultaneously by the three pairs of annotators on one text (Editorial CD 232323), which consists of 472 tokens. The initial inter-annotator agreement score was extremely low for this task at only 0.267, revealing a broad gap in agreement among the annotators, evidencing the high level of diversity that is expected for the pragmatic annotation of evaluative language.

Step 4: A second training session was carried out. The three annotation groups reviewed relevant aspects of Attitude and discussed the dis-agreements and problematic issues encountered during the annotation process. This training process eventually resulted in the formulation of a refined set of annotation guidelines.

Step 5: A second trial annotation was conducted on another text of 586 words (Editorial TG 20230223). The annotation this time resulted in a Fleiss kappa score of 0.812, demonstrating a significantly improved and satisfactory level of agreement.

Step 6: The groups proceeded to annotate the remaining corpus independently. The corpus was imported into the UAM Corpus Tool. Although the UAM Corpus Tool comes with some built-in layers for Appraisal Theory, we found it necessary to modify these layers to align with our specific annotation requirements. The resulting layers of the annotation scheme is illustrated in Figure 2. Text segments expressing emotional attitudes were manually identified and marked up through the selection of an appropriate tag.

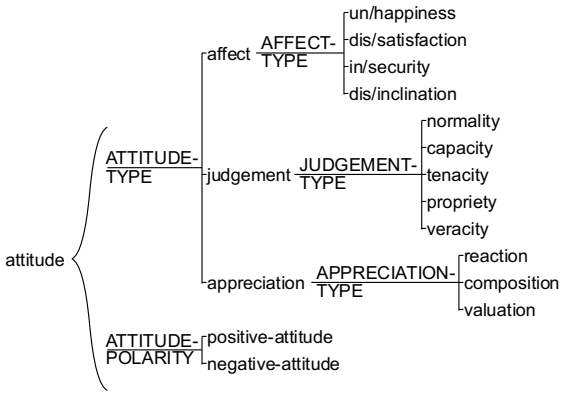

Figure 2: Refined annotation scheme of Attitude

3. Principles of Annotation

In what follows, we detail some of the major principles of annotation based on the two tests and outline the specific areas of disagreement encountered during the annotation process. We focus on the identification and categorization of evaluative language in a particular stretch of discourse, aiming to illustrate our practical methodological strategy to capture and classify evaluative expressions within texts with a good level of transparency and consistency.

3.1 Identifying What Needs to Be Annotated

The fundamental step in annotating evaluative language involves discerning which segments of text require annotation. Our principle is to identify and mark the smallest text segment that conveys the overall attitude or evaluative stance, which ensures precision and relevance in our annotations while capturing the attitude embedded within the text. Efforts were made to maintain a full phrase structure. Consider

(1-1) *Wang Huiyao says Beijing is **best** (+Valuation) placed to help negotiate an end to Russia's war in Ukraine.*

(1-2) *Wang Huiyao says Beijing **is best placed** (+Capacity) to help negotiate an end to Russia's war in Ukraine.*

In (1-1) and (1-2), we encounter possible annotation segments of "best" and "is best placed". While "best" alone might suggest a positive Valuation, annotating the broader phrase *is best placed* captures a more specific and contextually rich expression of positive Capacity.

3.2 Contextual Considerations in Annotation

The second principle extends beyond the identification of the smallest meaningful unit to encompass the contextual considerations of nouns that inherently express attitudes. Nouns such as "sanction", "conflict", and "invasion", while potentially evaluative, are approached with caution in specific contexts where they often serve a descriptive role, reflecting the factual dimensions of the situation rather than an evaluative stance. This principle acknowledges the importance of context in determining the evaluative nature of nouns.

3.3 Determining the Specific Category for Annotation

In categorizing annotated items, our approach is informed by principles outlined by Martin and White (2005) and further emphasized by Bednarek (2009). We aimed to differentiate between types of attitudinal lexis and evaluated targets or types of assessment. In practice, this means categorizing expressions related to emotions or feelings of people as Affect, evaluations of behaviour as Judgement, and assessments of objects or phenomena as Appreciation. This classification is instrumental in aligning evaluative expressions with the appropriate domain of appraisal, ensuring that our analysis is both systematic and aligned with the theoretical underpinnings of Appraisal Theory.

Once the primary category is determined, the next step involves specifying the subcategory based on the meaning. This process requires a careful analysis of the text to discern the specific nature of the evaluative stance being expressed. Our principle here emphasizes the importance of a detailed and context-sensitive approach to annotation. Bednarek's (2009) emphasis on the distinction between types of attitudinal lexis and evaluation targets serves as a crucial reminder of the depth and specificity required in annotating evaluative language, thereby enhancing the analytical precision.

In short, to ensure clarity and consistency, the following principles were applied: identifying the minimal meaningful textual segments for annotation, considering the context to accurately capture evaluative meanings, and categorizing annotations based on types of attitudinal lexis and evaluation targets.

4. Problems and Solutions for Annotating Appraisal

In the actual process of annotating evaluative language, despite having established a set of guiding principles, we still encountered several

problems related to identifying and classifying evaluative expressions. This situation underscores the gap between theories and practice, revealing areas that demand refinement, hence suggesting the importance of putting semantic annotation schemes to tests with authentic texts. This section outlines these problems and describes solutions.

4.1 Challenges in Identifying Appraisal and Possible Solutions

In (2) below, the phrase *seeks to* could be interpreted as expressing an inclination, a positive evaluative stance towards the action that follows.

(2-1) *This targeting of civilians reveals that Putin seeks not only to win. He* ***seeks to*** *(+ Inclination)* ***demoralize*** *(-Propriety).*

(2-2) *This targeting of civilians reveals that Putin seeks not only to win. He seeks to* ***demoralize*** *(-Propriety).*

However, the verb *demoralize*, which carries a negative connotation (negative Propriety), is the focal point of the evaluative stance in this context. The challenge here concerns whether to annotate *seeks to* for its positive inclination towards an action or to focus solely on the negative evaluative stance conveyed by *demoralize*. To address this issue, we opted not to annotate *seeks to* based on the principle that we should focus on primary evaluative meaning and avoid polarity conflicts. By prioritizing the annotation of "demoralize" for its negative Propriety, we ensure that the primary evaluative stance of the sentence is captured. This approach aligns with our principle of marking the smallest unit that conveys the overall attitude, emphasizing the importance of clarity in expressing evaluative meanings. Not annotating *seeks to* helps to avoid potential conflicts in evaluative polarity (positive vs. negative) that could arise from annotating both expressions. This decision ensures that our annotations remain coherent and focused on the most salient evaluative aspects of the text.

A second challenge encountered during the annotation process concerned the decision on how many segments should be annotated within a single sentence. This challenge is exemplified by (3) below.

(3-1) ***No country has as much diplomatic clout with Russia*** *(+Capacity)* ***while also having equally good ties with Ukraine as China*** *(+Capacity).*

(3-2) ***No country has as much diplomatic clout with Russia while also having equally good ties with Ukraine as China*** *(+Capacity).*

The example presents a comparative assessment of China's diplomatic ties with Russia and Ukraine, leading to a question: Should this be annotated as exhibiting one instance of Capacity that covers the entire comparative structure, or as two separate instances of Capacity for each of the diplomatic relationships mentioned? We eventually decided on a single annotation for the unified concept approach. The decision to annotate sentence (3-2) as one instance stems from the recognition that the sentence articulates a singular, overarching evaluative stance regarding China's diplomatic capabilities. The comparative structure of the sentence suggests a holistic evaluative judgement rather than two distinct evaluations. It reflects the integrated nature of the evaluative statement, where the two aspects of China's diplomatic relations are not isolated evaluations but interconnected to produce a singular assessment.

A further challenge concerns the appropriate scope for annotating evaluative meanings, especially when a single term might embody the evaluation, but its full implication becomes apparent only in a broader context. This challenge is illustrated by example (4).

(4-1) *Over the past decade, Russia had gradually* ***transformed*** *(+Capacity) itself from a marginal player in Asian affairs into a potential "third force" amid rising Sino-US rivalry.*

(4-2) *Over the past decade, Russia had gradually* ***transformed itself from a marginal player in Asian affairs into a potential "third force" amid rising Sino-US rivalry*** *(+Capacity).*

Here, *transformed* implies a significant and beneficial change, potentially warranting an annotation as Capacity on its own. However, the broader context provided by the complete phrase offers a more comprehensive understanding of Russia's change in status. To address this, we decided to annotate the entire phrase *transformed itself from a marginal player in Asian affairs into a potential "third force" amid rising Sino-US rivalry* as Capacity in (4-2). This decision is based on the understanding that the full evaluative impact of Russia's transformation is most accurately captured when considering the entire phrase. This approach allows for a more precise capture of the evaluative meaning, acknowledging that the significance of the transformation encompasses not just the act of change (*transformed*) but its direction and outcome (from *a marginal player* to *a potential "third force"*).

4.2 Challenges in Classifying Appraisal and Possible Solutions

We encountered significant challenges during the practical implementation of classification. These challenges primarily stem from the inherent subjectivity in distinguishing attitudes and the

vague boundaries between different evaluative categories. These factors frequently led to discrepancies among annotators, underscoring the need for a refined approach to ensure consistency and transparency in the annotation.

A major challenge is found in distinguishing between Affect and Appreciation, which is particularly pertinent when considering categories such as Security (a subcategory of Affect) versus Reaction (a subcategory of Appreciation). Consider

(5-1) *Putin's position, and perhaps his life, is **at risk** (-Security) if there is another big Ukrainian victory.*

(5-2) *Putin's position, and perhaps his life, is **at risk** (-Reaction) if there is another big Ukrainian victory.*

It could be argued that the phrase *at risk* should be classified under Affect, focusing on Security as it highlights concerns for Putin's personal safety and political stability. This interpretation emphasizes the emotional impact and the sense of threat to well-being, suggesting Affect as the fitting category. Alternatively, the same phrase could be analyzed as Appreciation with an emphasis on Reaction. This analysis assesses the sentence as evaluating the consequences or outcomes of a potential event on Putin's position, considering it an evaluation of situational change rather than an emotional response. The choice between Affect and Appreciation thus hinges on the interpretation of the sentence's core focus. If viewed primarily as eliciting an emotional response regarding Putin's precarious situation, Affect is deemed appropriate. However, if the sentence is interpreted as assessing the impact of potential events on Putin's status, Appreciation would be chosen.

The differentiation between Affect and Judgement presents another layer of complexity in the annotation process, especially when sentences can potentially align with either category based on their evaluative focus. This challenge is illuminated in (6).

(6-1) *Ukrainians **have needlessly suffered a terrible toll** (-Happiness) and the impact is rippling around the world, with disruptions to food supplies and higher energy and grain costs bringing hunger and poverty to tens of millions of vulnerable people.*

(6-2) *Ukrainians **have needlessly suffered a terrible toll** (-Propriety) and the impact is rippling around the world, with disruptions to food supplies and higher energy and grain costs bringing hunger and poverty to tens of millions of vulnerable people.*

Opting to annotate as "-Happiness" suggests an interpretation focused on the emotional response elicited by the Ukrainians' suffering, reflecting the emotional distress and negative states, hence fitting the Affect category. Alternatively, a perspective on Propriety shifts the perspective towards a moral or ethical Judgement. This view interprets it as a violation of moral standards, emphasizing the situation's ethical implications over its emotional impact.

The ambiguities between Affect vs. Judgement and Affect vs. Appreciation are a notable challenge that has been identified in the literature. Thompson (2014) has referred to this as the "Russian doll effect", where evaluative meanings are nested within one another, potentially qualifying for multiple categories of appraisal. Double annotation has been advocated to capture the layered nuances of evaluative language (Macken-Horarik and Isaac, 2014). However, for the sake of consistency and simplicity in the annotation process, we decided to adhere to a single annotation and to categorize based on the most prominent aspects: Affect for evaluations relating to emotions or feelings of people, Judgement for behaviours or actions, and Appreciation for objects or phenomena.

A single annotation approach simplifies the process, making it more accessible and manageable for annotators. Double annotation, while potentially offering a richer analysis, introduces complexity that could hinder the efficiency and consistency of the annotation process. The single annotation can be supplemented by a comprehensive textual analysis at a later stage, which will allow for a deeper exploration of the texts, where the nuances that might have been simplified during the annotation can be revisited and analyzed in greater depth. This strategy does not overlook the complexity of evaluative language but rather postpones a more granular analysis to the post-annotation stage. Here, the annotations serve as a foundation for further reflection and investigation, allowing researchers to explore the "Russian doll effect" with the full context of the text in view. This reflective analysis enables us to understand how evaluative meanings are interwoven and how they contribute to the overall discourse.

We have discussed the primary distinctions among Affect, Judgement and Appreciation. These distinctions are critical for identifying the broad categories in which language can express evaluations and attitudes. Finer distinctions need to be investigated, especially in Judgement. Consider

(7-1) *The cold shoulder: Richard Heydarian says the Ukraine invasion **has soured Russia's ties across Southeast Asia** (-Normality).*

(7-2) *The cold shoulder: Richard Heydarian says the Ukraine invasion **has soured Russia's ties across Southeast Asia** (-Propriety).*

(7-1) is labelled as -Normality, suggesting that the Ukraine invasion is being evaluated in terms of its deviation from expected or conventional diplomatic behaviour, thus affecting Russia's international relationships. The focus is on the abnormality of the situation, implying that such actions are not in line with what is typically expected in international relations, leading to a deterioration in ties. Alternatively, it can also be annotated as negative Propriety in (7-2), shifting the emphasis to the appropriateness of the invasion and its consequences. This perspective assesses the invasion's impact on diplomatic relationships as a matter of ethical judgement, suggesting that the action is morally wrong or unacceptable, hence the negative repercussions on Russia's relations. Given the nuanced differences between Normality and Propriety within the Judgement category, where Normality is associated with social esteem and Propriety with social sanction, the challenge arises in ensuring accurate and consistent annotation.

To address this challenge and enhance both inter-rater agreement and consistency, we decided to prioritize Propriety when overlapping occurs. When an evaluative statement could potentially be annotated as both Propriety and Normality, the guidelines should advise annotators to prioritize Propriety. The prioritization is grounded in the intrinsic relationship and hierarchy between these concepts. Propriety encompasses appropriate-ness, which inherently requires actions or behaviour to align with societal norms and expectations, thus implying Normality. However, Normality focuses solely on the conformity of actions with norms and standards without necessarily engaging with their moral or ethical dimensions. Propriety assessments include a judgement of Normality but also extend beyond to consider legal appropriateness. By adopting Propriety as the default category in cases of overlap, annotators are likely to achieve higher consistency in their evaluations.

Distinguishing between Capacity and Tenacity within the Judgement category presents another layer of complexity. Both subcategories pertain to evaluations of behavior, but they focus on different aspects. Capacity refers to the ability or power to do something, often related to skill or competence. Tenacity, on the other hand, emphasizes persistence or determination in pursuing goals, especially in the face of obstacles.

(8-1) *Through its permanent seat on the United Nations Security Council, it also has the means to **ensure that countries adhere to global standards** (+Capacity).*

(8-2) *Through its permanent seat on the United Nations Security Council, it also has the means to **ensure that countries adhere to global standards** (+Tenacity).*

To navigate the distinction between Capacity and Tenacity more effectively in (8), we have to carefully examine the context to identify whether the emphasis is on the inherent ability (Capacity) or on the persistence and determination (Tenacity). Tenacity often implies a sustained effort in the face of challenges or obstacles. If the text highlights overcoming difficulties or persistent effort, Tenacity might be the more appropriate category. In contrast, Capacity focuses on the ability or competence without necessarily implying effort against resistance.

An additional issue is the disproportionate representation of the Valuation subcategory within Appreciation compared to Reaction and Composition. As illustrated in Figure 3, the instances of Valuation across the four newspapers significantly outnumber those of the other two subcategories within Appreciation. It emerged during the annotation that when segments did not clearly align with Reaction or Composition, there was a tendency to categorize them as Valuation.

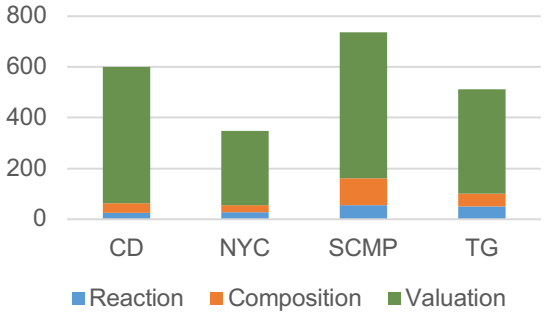

Figure 3: Distribution of subcategories under Appreciation for different newspapers

The use of Valuation as a catch-all category, while streamlining the annotation process, introduced challenges in analysis. The over-representation of Valuation could dilute the specificity of our findings, making it harder to discern distinct patterns or nuances in evaluative expressions. Such a broad categorization risks oversimplifying the rich evaluative landscape present in discourse, potentially masking the intricate ways in which objects or phenomena are appraised. To address this issue, it is crucial to refine the criteria for categorizing evaluative expressions under Valuation. This refinement process may involve expanding the Appreciation dimension to include additional, more specific categories tailored to the texts being analyzed. By doing so, we can accommodate a broader range of evaluative expressions, ensuring a more granular and accurate classification. Thus, while annotating, there is an opportunity to extend the Appreciation

categories as needed, ensuring that the framework remains flexible and responsive to the complexities of the texts under examination.

5. Conclusion

Throughout this project, we explored the practical application of Appraisal Theory in the task of corpus annotation with a particular focus on the Attitude system. The endeavour was driven by the aim to illuminate the complexities and challenges inherent in the annotation process and to come up with effective strategies for overcoming these obstacles. Central to our project was the formulation and implementation of a set of annotation guidelines to ensure accuracy and consistency. These principles guided our approach to identifying evaluative expressions, considering their contextual implications and categorizing them accordingly. Through this practical methodology, we aimed to refine the process of corpus annotation, making it a more effective tool for semantic annotation in general and pragmatic annotation of stance in particular.

Our annotation experiment revealed significant issues, particularly in the dual tasks of identifying and classifying evaluative expressions within the texts, highlighting the complexity of Appraisal Theory and the inherent subjectivity in interpreting expressions of Attitude. Our corpus-informed solutions involved a detailed examination of the Attitude category in authentic texts, leading to the formulation of strategies to resolve specific confusable annotations. This approach facilitated a more structured annotation process contributing to the broader issue of semantic and pragmatic annotation of corpus data involving subjective judgements. Moreover, this study identified a need for flexibility within the annotation framework, especially in addressing the disproportionate use of the valuation subcategory within Appreciation. This observation prompted a critical re-evaluation of our classification strategy, allowing for a more granular analysis of evaluative language.

During the refinement of our annotation guidelines within the Appraisal Theory framework, we realized that incorporating parts of speech (POS) and phrasal structures into our definitions of annotation units had not been explicitly stated. Addressing this could substantially enhance the degree of transparency and consistency in the identification of evaluative segments. Insights from Caro (2014) and Hunston and Su (2019), who emphasize the evaluative potential of adjectives, nouns and verbs, suggest that future efforts could adopt a hierarchical approach to annotation. Such an approach would give precedence to adjective phrases due to their prominent role in conveying evaluative meaning while still recognizing the contributions of nouns and verbs. Additionally, it was decided that nouns derived from adjectives and verbs should be annotated accordingly. Our updated strategy for selecting the smallest text segment for annotation advocates a flexible method: starting with single lexical items, then expanding to phrases, and eventually to clauses if necessary. Future enhancements to our annotation guidelines might also benefit from including phrasal and syntactic structures. Acknowledging the syntactic roles of adjectives, nouns, and verbs within their respective phrases could help more precisely to identify the scope of evaluative expressions. It should be noted that while the discussions so far have centred on grammatical aspect, semantic factors are fundamentally important and form a major basis of annotation judgements.

In conclusion, our experiment reported and addressed the practical task of annotating pragmatic information within the framework of Attitude in Appraisal Theory. It has detailed the process of identifying and classifying evaluative expressions, thereby enhancing both transparency and consistency in the annotation practices. By proposing specific solutions to the intricacies involved in the annotation of evaluative language, this work contributed towards the methodological foundations for future research. Our future work will incorporate parts of speech and phrasal structures into annotation guidelines and adopt a hierarchical approach to better capture evaluative text segments. We plan to integrate parts of speech and phrasal structures into our annotation guidelines, employing a hierarchical approach to identify evaluative text segments more consistently in conjunction with semantic considerations.

6. Acknowledgement

This work was supported in part by grants received from China's National Planning Office of Philosophy and Social Sciences (Project No 22BYY009), Beijing Social Sciences Foundation (Project No. 18JDYYA005) and City University of Hong Kong (Project Grant Nos 7020036, 9360115 and 6008167). The second author would like to thank the Halliday Centre for Intelligent Applications of Language Studies at City University of Hong Kong for a visiting professorship received in 2023.

7. Bibliographical References

Bednarek, M. (2006). *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. Continuum.

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

Bednarek, M., & Caple, H. (2010). Playing with environmental stories in the news – Good or bad practice? *Discourse & Communication, 4*, 5–31.

Breit, B. W. (2014). Appraisal theory applied to the wine tasting sheet in English and Spanish. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos*, (27), 97-120.

Caro, E. M. (2014). The expression of evaluation in weekly news magazines in English. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context* (pp. 321–343). Amsterdam and Philadelphia: 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.

Geng, Y., & Wharton, S. (2016). Evaluative language in discussion sections of doctoral theses: Similarities and differences between L1 Chinese and L1 English writers. *Journal of English for Academic Purposes, 22*, 80-91.

Hood, S. (2010). *Appraising research: Evaluation in academic writing*. Springer.

Huan, C. (2016). Journalistic engagement patterns and power relations: Corpus evidence from Chinese and Australian hard news reporting. *Discourse & Communication, 10*(2), 137-156.

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

Macken-Horarik, M., & Isaac, A. (2014). Appraising appraisal. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context* (pp. 67–92). Amsterdam and Philadelphia: John Benjamins.

Martin, J. R. (2000). Beyond exchange: Appraisal systems in English. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 142–175). Oxford: Oxford University Press.

Martin, J. R. and White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English.* London: Palgrave.

Mayo, M. A., & Taboada, M. (2017). Evaluation in political discourse addressed to women: Appraisal analysis of Cosmopolitan's online coverage of the 2014 US midterm elections. *Discourse, context & media, 18*, 40-48.

Meadows, B., & Sayer, P. (2013). The Mexican sports car controversy: An appraisal analysis of BBC's Top Gear and the reproduction of nationalism and racism through humor. *Discourse, Context & Media, 2*(2), 103-110.

O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pp. 13–16. Association for Computational Linguistics.

Pounds, G. (2011). This property offers much character and charm: Evaluation in the discourse of online property advertising. *Text & Talk, 31*(2), 195–220.

Swain, E. (2010). *Getting engaged: Dialogistic positioning in novice academic discussion writing*. EUT Edizioni Università di Trieste.

Taboada, M., Carretero, M., & Hinnell, J. (2014). Loving and hating the movies in English, German and Spanish. *Languages in Contrast, 14*(1), 127-161.

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