From Theory of Rhetoric to the Practice of Language Use: The Case of Appeals to Ethos Elements

Marcin Koszowy¹ · Katarzyna Budzyska¹ · Martín Pereira-Fariña² · Rory Duthie¹

Accepted: 9 December 2021 / Published online: 3 January 2022
© The Author(s) 2021

Abstract

In their book *Commitment in Dialogue*, Walton and Krabbe claim that formal dialogue systems for conversational argumentation are “not very realistic and not easy to apply”. This difficulty may make argumentation theory less well adapted to be employed to describe or analyse actual argumentation practice. On the other hand, the empirical study of real-life arguments may miss or ignore insights of more than the two millennia of the development of philosophy of language, rhetoric, and argumentation theory. In this paper, we propose a novel methodology for adapting such theories to serve as applicable tools in the study of argumentation phenomena. Our approach is both *theoretically-informed* and *empirically-grounded* in large-scale corpus analysis. The area of interest are appeals to ethos, the character of the speaker, building upon Aristotle’s rhetoric. Ethotic techniques are used to influence the hearers through the communication, where speakers might establish, but also emphasise, weaken or undermine their own or others’ credibility and trustworthiness. Specifically, we apply our method to Aristotelian theory of ethos elements which identifies *practical wisdom*, *moral virtue* and *goodwill* as components of speakers’ character, which can be supported or attacked. The challenges we identified in this case and the solutions we proposed allow us to formulate general guidelines of how to exploit rich theoretical frameworks to the analysis of the practice of language use.

Keywords Theory of argumentation vs practice of argumentation · Bottom-up methodology · Corpus analysis · Aristotle’s rhetoric · Ethos · Ethos elements · Wisdom—Virtue—Goodwill

✉ Marcin Koszowy
marcin.koszowy@pw.edu.pl

¹ Laboratory of The New Ethos, Warsaw University of Technology, Warsaw, Poland

² Department of Philosophy and Anthropology, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

1 Introduction

Despite attempts at employing rich philosophical and rhetorical theories as conceptual frameworks to capture everyday argumentation, the robust methodology, which would systematically link those theories with the practice of language use, is still lacking. As some research in formal dialogue systems and argument mining has shown, the conceptual apparatus of the theories, as it stands, is usually difficult to apply in the study of pragmatic features of argumentation. Since there is no general methodological solution that would help to step-by-step adapt theories to natural language data, in this paper we propose and deploy a novel methodology that links theory and practice of argumentation in the area of appeals to ethos—an area of special importance, as it tackles one of the most effective rhetorical tools.

Among various kinds of evidence for the lack of robust methods to bridge theories with data, there are two examples related, respectively, to the application of a theory of formal dialogue systems to evaluating natural dialogues (Walton and Krabbe 1995), and to the application of a theory of argumentation schemes to annotating ethos for argument mining [(Habernal et al. 2018) (see also (Hinton 2021))] on the role and importance of argument corpora for studying argumentation]. In the first case, the major problem for the logical evaluation of conversational argumentation has been “the rigor and precision of a system of logic, on the one hand, and the permissive, free flow of ordinary conversation on the other” (Walton and Krabbe 1995, p. 174). The typical solution to this problem was to employ formal dialogue systems such as those proposed by Lorenz and Lorenzen (1978), and Barth and Krabbe (1982), yet, as Walton and Krabbe stated, such systems are “not very realistic and not easy to apply” (*ibid.*).

Some studies in the area of argument mining have also proven that theories are usually not directly applicable to the study of complex linguistic data. Habernal et al. (2018) used existing theories of ad hominem arguments for the purpose of mining them from natural language data. In their annotation guidelines, the labels for five types of ad hominem: abusive, *tu quoque*, circumstantial, bias and guilt by association, were specified according to standard accounts available in the literature. The results of annotation turned out to be disappointing with a dominance of one type (41% out of 200 instances of ad hominem were analysed as abusive) and significant disagreements between annotators with respect to the rest of ad hominem types. The authors concluded that “(1) the theoretical typology does not account for longer ad hominem arguments that mix up different attacks and that (2) there are actual phenomena in ad hominem arguments not covered by theoretical categories”. This means that such a typology and an annotation cannot be applied to build a technology for argument mining, and the authors decided to build their own typology from scratch which completely ignored any insights developed in theories.

These are just two examples of a gap between theoretical models of conversational argumentation and studies which require models that are empirically grounded in the practice of language use, but the same will hold, in particular,

for any *linguistic study* of conversational argumentation, i.e. when annotation of argumentation and ethos is performed manually, such as e.g. in Goodwin and Cortes (2010), Wagemans (2011), Wacholder et al. (2014), Musi and Aakhuis (2019), Pereira-Fariña et al. (2022), as well as for its *computational study*, i.e. when we aim to automatically mine arguments (cf. Stede and Schneider 2018; Lawrence and Reed 2020; Visser et al. 2021), or we aim to mine ethos (cf. Habernal et al. 2018; Duthie and Budzynska 2018a), or to mine both (cf. Musi et al. 2016; Hidey et al. 2017; Wachsmuth et al. 2018; El Baff et al. 2019). This shows that in order to be able to apply a theoretical account for the empirical study of conversational argumentation, a new methodology is needed.

The task of proposing such a methodology will be elaborated in this paper on the case of a fairly complex and important communication phenomenon, namely appeals to ethos, or more specifically - appeals to ethos elements. Following Aristotle's theory of rhetoric, we investigate ways in which speakers appeal to three ethos elements: practical wisdom (*phronêsis*), moral virtue (*aretê*) and goodwill (*eunoia*). The area of appeals to ethos elements turned out to be a paradigmatic example of challenges which arise when we import rhetorical and philosophical theories to the large-scale analysis of the practice of language use.

As a resource which contains a noticeable number of appeals to ethos, we take Hansard, the UK parliamentary debate record. In our method of linguistic analysis of ethos elements, we build upon the approach presented in Duthie and Budzynska (2018a) where Aristotle's definition of ethos (Aristotle 1991, II.1) was employed to specify ethotic expressions that target properties of the individual or the group of politicians. Within this framework, a sentence uttered by a source-speaker targets ethos of a referent-speaker who can be either supported (Miss Widdecome in Example (1-a)) or attacked (the British Government in Example (1-b)):

- (1) a. Mr. John Moore: *I bow to my hon. Friend's [Miss Widdecome] distinguished past and detailed knowledge of these matters.*
- b. Mr. Bruce Grocott: *Is it not the simple truth that the Government are making the country sick?*

This work was further extended in (Duthie and Budzynska 2018b) to automatically mine more-fine grained types of ethos supports and attacks in which different grounds for these supports and attacks can be distinguished: in Example (1-a) Mr. Moore is supporting Miss Widdecome because of distinguished past and detailed knowledge, while in (1-b) Mr. Grocott is attacking the Government, because they have made the country sick (i.e. not acting in the country's best interest). This corresponds to the Aristotelian distinction between ethos elements (Aristotle 1991, II.1, 1378a6ff): practical wisdom, moral virtue and goodwill (see Sect. 2.1). Thus, (1-a) can be categorised as an instance of grounding the support in the referent-speaker's wisdom, while (1-b) as an instance of grounding the attack in the referent-speaker's goodwill.

As a result of employing Aristotelian ethos elements as a conceptual framework to label ethotic expressions, we introduced six basic categories which we use

to annotate natural argumentation in parliamentary discourse, including: (i) three types of ethos supports from: Wisdom (W^+), Virtue (V^+) and Goodwill (G^+); and (ii) three respective types of ethos attacks (W^- , V^- , G^-). The core of the proposed methodology is an iterative process of refining the theoretical framework of Aristotle's ethos theory step-by-step to make it able to grasp linguistic features of ethotic expressions in natural argumentation. As a result, three iterations of annotating ethos elements were designed in the spirit of agile corpus creation (Voormann and Gut 2008) allowing us to gradually adapt the theory to the practice of argumentation. For instance, as the second iteration of ethos elements annotation has revealed the complex and unresolvable overlaps between those basic categories, in the next iteration we enriched the annotation scheme by introducing overlapping categories, called polymorphic ethos elements (e.g. WV^+ , WG^+ , VG^+ and WVG^+) to capture those instances of ethotic expressions that can be justifiably located on the overlaps between the three basic categories.

We first discuss the related work (Sect. 2) and the background for our research methodology (Sect. 3). Then, using the case of Aristotle's rhetorical account of ethos elements, we propose a novel bottom-up methodology which refines a theoretical model into an empirically-grounded model that is then applicable to appeals to ethos elements occurring in conversational argumentation (Sect. 4). Finally, we formulate general suggestions which can further guide the adaptation and application of rich theoretical frameworks to the analysis of the practice of language use (Sect. 5).

2 Related Work

Our methodology will be applied to study ethos on the assumption of its indispensable and ubiquitous role in persuasive communication recognised by rhetoric (Baumlin and Scisco 2018). Some advances in argumentation theory either point to the need of a systematic study of ethotic arguments (Brinton 1986; Walton 1999; Tindale 2011) or emphasise the role of trust (Liao 2021) and distrust (Jackson 2015) as ways through which ethos is employed in communication. In this section, we set the grounds of our approach, which involves the Aristotelian theory of ethos as an inspiration for ethos analysis (Sect. 2.1), and contemporary related work (Sect. 2.2).

2.1 Aristotelian Theory of Ethos

In his *Rhetoric*, Aristotle investigated ethos, the character of the speaker as one of the three modes of persuasion together with logos, the structure of the argument itself, and pathos, the emotional state of the listener. Specifically, ethos is the proponent's character revealed through communication. Aristotle emphasises speakers' dispositions as a certain kind of persons to be a key element of successful persuasion so that "their hearers suppose them to be disposed toward them in a certain way" (Aristotle 1991, II.1, 1377b25ff). Aristotle recognises the special role of ethos in communication, as it "may almost be called the most effective means of persuasion

he [a proponent] possesses” (Aristotle 1991, I.2,1356a13), in particular when the issue discussed is difficult, dubious or uncertain. If the audience is unable to fully understand and follow arguments which prove that the issue is true (or probable), then it is morally justified for the speaker to partly use ethos as a substitute for logos.

Three elements of ethos, i.e. three elements of speaker’s character, can create hearers’ favourable or unfavourable disposition toward him: practical wisdom (gr. *phronêsis*), moral virtue (gr. *aretê*), and goodwill (gr. *eunoia*). Aristotle points to them as to “three reasons why speakers themselves are persuasive”, because “through lack of practical sense they do not form opinions rightly” (the lack of practical wisdom), “though forming opinions rightly they do not say what they think because of bad character” (the lack of moral virtue), or “they are prudent and fair-minded but lack good will, so that it is possible for people not to give the best advice though they know [what] it [is]” (Aristotle 1991, II.1, 1378a6ff).

The hearers perceive the speaker as *wise*, when they assume that they know the truth with regard to the subject of the speech. Practical wisdom is rooted in a speaker’s experience on a matter in which they advise. For example, a general who proved to be a good strategist may be considered as a wise speaker, if he advises about tactics that would win a war. The audience treats the speaker as *virtuous*, when they believe the speaker says what they really think, i.e. when the speaker does not lie. If the virtuous speaker advises a democratic city, they should possess the virtue of the good democrat. Aristotle also associates the virtue with being immune to any temptations like bribery or flattery. If the speaker shares an information which they do not believe to be true, then their ethos may be impaired. When the speaker demonstrates *goodwill*, then the audience will assume that the speaker shares the truth with them, as long as it is known. This ethos element can be associated with the trait of being friendly, characterised as “wishing-well” for others for their own sake (Aristotle 1991, II.4, 1380b–1381a). Thus, in the general sense, goodwill can be treated as showing an alignment between a speaker’s own interests and interests of the audience.

If the speaker is only wise, the audience may doubt whether the speakers aims are good, and if the speaker is only wise and virtuous, then the audience may doubt whether the speaker gives the best advice they could (Rapp 2010). Thus, the speaker needs to possess (or be seen as possessing) all three ethos elements in order to be persuasive. Moreover, a preexisting good character is not necessarily known or considered by the audience, hence the linguistic expression of ethos is an essential part of the discourse. Hence, the Aristotelian account of ethos is a firm motivation for studying ethotic language in order to capture the diversity of appealing to ethos elements within a unified theoretically-informed and empirically-grounded approach. In Sect. 2.2, we further explore related work by discussing the place of ethos elements in contemporary argument studies.

2.2 Argumentation Theory on Ethos

The Aristotelian account of ethos has been incorporated into argumentation theory to investigate argumentation schemes—stereotypical patterns of reasoning (*cf.*

Walton et al. 2008). In *ethotic argumentation* (Brinton 1986) or *interpersonal argumentation* (Budzynska 2010), these schemes determine that a statement or argument is true if the author of that statement or argument is a person of good character, consequently, a statement or argument should not be accepted if the author of that statement or argument is a person of bad character (cf. Walton et al. 2008, pp. 140–141). In the first case, in so called positive ethotic arguments or pro homine arguments, the conclusion “A should be accepted” is inferred from premises that “X said A” and “X is knowledgeable, trustworthy, and free of bias” (Groarke and Tindale 2013, pp. 368–369). Negative ethotic arguments or ad hominem arguments are conceived as counter-arguments to pro-homine arguments: “good ad hominem arguments usually appear in contexts where an appeal to a pro homine has occurred or might occur” (Groarke and Tindale 2013, pp. 378–379). As Aristotelian ethos elements can be identified in both positive and negative forms of ethotic argumentation studied in the literature, in what follows we will briefly make those affinities plain.

In the case of positive ethotic arguments, the ethos element most frequently exposed in the current takes of ethotic arguments is that of wisdom. That is due to the fact that typical positive ethotic arguments such as arguments from expert opinion and arguments from position to know strongly emphasise the epistemic component of ethos that rests upon knowledge and skills (Walton 1997; Goodwin 2011; Mizrahi 2013; Seidel 2014; Koszowy and Walton 2017). For instance, argument from expert opinion (Walton 1997, p. 210) consists of drawing a conclusion from a premise which states the speaker’s expertise in a given domain. It should be stressed here that there is a clear distinction between practical and theoretical wisdom. In Aristotle’s work, the focus was on practical knowledge and practical experience, while modern argumentation theory focuses on the theoretical wisdom of experts which is similar to the approach in epistemology concerned with scientific and legal expertise (cf. Hardwig 1991; Goldman 2001). The only exception is the recent study on arguments from deontic authority in which we reason about what should be done from authority of a speaker who is authorised to give orders or advices (cf. Araszkiewicz and Koszowy 2016; Koszowy and Walton 2019).

Moreover, in the procedure of assessing arguments from expert opinion, some of the critical questions aim at verifying experts’ possible bias related to expert’s character for veracity or experts’ possible honesty related to whether or not s/he has lied in the past, has a criminal record or there are “reasons to think that an expert is lying or is not being sincere in advancing the opinion in question” (Walton 1997, p. 217). Asking critical questions about those ethotic features is clearly a reference to moral virtue as a distinct ethotic trait.

In the case of negative ethotic arguments, the presence of ethos elements can be traced with respect to the particular types such as abusive, circumstantial, guilt by association and poisoning the well (Macagno 2013). An abusive ad hominem technique (Hansen 2019) may be associated with attacking each ethos element: practical wisdom, when opponents are claimed to be ignorant; moral virtue, while stating that the other party is morally bad, malicious etc.; and goodwill, when the other party is, e.g., claimed not to be honest with the audience by misleading them. In circumstantial ad hominem (Wrisley 2019), an attack can point to the bias of a source on the grounds of: practical wisdom, when what the other party claimed at some

point is in conflict with what that party claimed at another time; moral virtue, when the values claimed to be held by a speaker are claimed to be in conflict with their morally bad attitude or action; or goodwill, where the declarations of someone's aligning with the audience are claimed to be in conflict with how that person really verbally behaves towards the audience. Guilt by association is potentially related to also attacking all three ethos elements, because an attacked person can be claimed to be associated with a group which is assessed as a group of ignorants (practical wisdom), a group of people of morally bad character (moral virtue) or a speakers group that does not want to have any kind of alignment with their audience (goodwill). Finally, poisoning the well is mostly related to attacking practical wisdom: it is considered an *ad hominem* as the adverse information typically has no relevance towards the logic of a person's argument.

The discussed presence of ethos elements in particular types of ethotic arguments points to the need for a systematic inquiry aimed at collecting evidence about whether practical wisdom, moral virtue and goodwill are in fact typical in an argumentative discourse. In Sect. 3, we present the foundations for a bottom-up methodology that will be then applied to adapting Aristotle's account of ethos elements to the study of linguistic data.

3 Background

This section first describes textual material of parliamentary debates and the original annotation of *ethos supports* and *ethos attacks* in these data (Sect. 3.1). Then, we show how this data can be re-annotated into ethos supports and attacks *from practical wisdom, moral virtue and goodwill* (Sect. 3.2). Finally, we present the method of applying the re-annotation in an iterative process which is in fact the core idea constituting the proposed method 3.3.

3.1 Annotation of Ethos

In this paper, we build on the resources developed in Duthie and Budzynska (2018a) where the simplest structures of *ethotic expressions* were investigated, i.e. *ethos supports* and *ethos attacks*. This work is founded upon the meaning of ethos from Aristotle (1991) and then extends it to account for appeals to the character of politicians and groups of politicians. As a textual material for the study of the language of ethos elements, this work selects the UK parliamentary debate record, *Hansard*, which stores all the sessions in the British government dating back to 1800 (available at <http://hansard.millbanksystems.com>). The transcripts analysed were taken from the late period of Margaret Thatcher's government (from 1979 to 1990) which has been a relatively dynamic and controversial period in British politics, thus expecting a large number of appeals to character. The dataset covers transcripts of 90 parliamentary sessions, with a total of 90,990 words.

This work aimed to develop *Natural Language Processing* technology of ethos mining to automatically extract the relationships between politicians or between

Table 1 Summary of the *EthosHansard2* corpus: categories and labels used for annotating ethos, and inter-annotator agreement (kappa)

Categories Labels	Speakers		Ethotic expressions	
	Source	Referent	Ethos supports	Ethos attacks
#	198		638	
	149	188	169	469
%	75%	95%	26%	74%
Kappa κ	–		0.67	
	1	0.93	1	1

a politician and a party through ethotic sentiment expressions. Ethos is specified here as a property of the individual or the group of agents which can be supported or attacked in order to influence the audience through communication. When a politician (source-speaker) supports the ethos of another politician or party (referent-speaker), then a positive sentiment expression is identified (Examples (2-a)–(2-c)); while attacks on ethos are classified as negative sentiment expressions (Example (2-d)). The annotation uses cues of the polarity of sentiment signalled by source-speakers on the linguistic surface such as ‘distinguished past’ and ‘detailed knowledge’ in Example (2-a), and ‘sick’ in Example (2-d) (all linguistic cues are in bold; full guidelines are available at <https://arg.tech/EthosHansard2-Guide>).

- (2) a. Mr. John Moore: *I bow to my hon. Friend’s [Miss Widdecome] **distinguished past** and **detailed knowledge** of these matters.*
- b. Mr. Patrick Jenkin: *I believe that the Government were right to **have the courage** to bring forward the necessary measures to bring public expenditure under control.*
- c. Mr. John Moore: *My hon. Friend [Sir Anthony Meyer] is assiduously **pursuing his constituents’ interests***
- d. Mr. Bruce Grocott: *Is it not the simple truth that the Government are making the country **sick**?*

The annotation of ethotic expressions resulted in the *EthosHansard2* corpus (available at <http://corpora.aifdb.org/EthosHansard2>). The dataset contains 198 unique speakers (see Table 1) of whom 95% were targets of supports or attacks, and 75% were the authors of supports and attacks. Ethos supports were significantly less frequent than attacks (26% vs 74%) which is the reversed proportion observed for argument supports and argument attacks. The reliability of the annotation was measured with Inter-Annotator Agreement which involves a comparison between two or more annotators. For this task Cohen’s kappa (Cohen 1960), a statistic which takes into account annotators agreement by chance, was measured between two annotators on a randomly selected 10% subset of the data. The reliability of annotation of speakers is perfect or almost perfect for source speakers and referent speakers (kappa of 1 and 0.93, respectively). The reliability of ethotic expressions at $\kappa = 0.67$ is considered to be substantial (Landis and Koch 1977), while once an expression

Fig. 1 Ethos support from wisdom in Example (2-a) (labelled as W^+)

is determined to be ethotic, the agreement of deciding whether it is a support or an attack is perfect ($\kappa = 1$).

3.2 Annotation of Ethos Elements

In Duthie and Budzynska (2018b), this resource has been further re-annotated to account for different grounds for ethos attacks and supports: the politician's practical wisdom, moral virtue and goodwill. Whereas Example (2-a) illustrates Mr. Moore's support of Miss Widdecome's wisdom (signalled by 'distinguished past' and 'detailed knowledge'), the appeal to ethos made by Mr. Jenkin (2-b) targets virtue ('have the courage'), and the appeal by Mr. Moore (2-c) contains a reference to the politician's goodwill ('pursuing his constituents' interests'). Apart from ethos supports (or ethotic premises) there are also ethos attacks (or ethotic conflicts), one of which is illustrated in Example (2-d).

The annotation and re-annotation has been done with the use of the OVA+ software tool (Janier et al. 2014) (<http://ova.arg-tech.org/>) which allows the analysts to reveal the underlying language structures of ethotic expressions. It builds upon Inference Anchoring Theory (IAT) (Budzynska and Reed 2011; Budzynska et al. 2016) which describes arguments and their dialogical context including arguments which appeal to character. The key advantage of this framework is that IAT assumes that argumentation is a pragmatic phenomenon, which allows us to capture how *language is used* in actual communication practice. The result of the annotation is stored in the freely available AIFdb repository (Lawrence et al. 2015) (<http://aifdb.org>) which allows us to create corpora for subsequent iterations of annotation and to make them publicly available.

IAT makes use of the distinction between an utterance performed in the discourse (on the right hand side of the diagram such as in Fig. 1) and its propositional content (on the left hand side of a diagram). Figure 1 shows the annotation of ethos support where the "X has ethos" node is supported through *Argument from Practical Wisdom* by a propositional content of what John Moore has said during the parliamentary session. Re-annotation of ethos with ethos elements means that the general label of ethos support, i.e. *Default Inference*, is replaced by a specific fine-grained

Fig. 2 Ethos Attack against Goodwill in Example (2-d) (labelled as G^-)

type of ethos support, i.e. *Argument from Practical Wisdom*, which can be then again replaced by another label in a subsequent iteration of annotation. In the same style, Fig. 2 shows the annotation of an attack on the ethos element of goodwill: the propositional content of Bruce Grocott’s assertive question attacks the Government’s ethos through the *Conflict from Goodwill* node which replaced an original annotation of ethos (annotated with the *Default Conflict* node). We will refer to this type of appeal as “attack against goodwill” or “in conflict with goodwill”, but we keep the notation *Conflict from Goodwill* in annotated maps, following the terminological convention used in the OVA+ software tool.

The work in Duthie and Budzyska (2018b) focused heavily on the development of the technology of ethos mining, thus it treated annotation of ethos elements as rather marginal. As a result, it used an original account of Aristotelian ethos elements to re-annotate the corpus described in the previous section. As the reliability of this annotation turned out to be too low for the purpose of automation, this basic annotation scheme was then improved to fix the main errors encountered in the manual annotation. These two steps of annotation will constitute the foundations for this paper for the first and second iterations of annotation described in Sects. 4.2 and 4.3. Yet, the error analysis of the first iteration has been neither discussed from the perspective of its consequences for the model of ethos elements nor even briefly presented in the paper. In this sense, this work complements (Duthie and Budzyska 2018b) by describing the theoretical and methodological implications of annotating textual data. Moreover, it extends the previous work by running a detailed error analysis of the second iteration of annotation which results in a significantly more complex annotation guidelines for ethos elements. This error analysis and annotation guidelines provide a key insight into the process of moving from a theory of rhetoric to the practice of language use.

3.3 Agile Corpus Creation

Our methodology is based on a process used for annotating corpora, the agile corpus creation (Voormann and Gut 2008), which consists in cycles or iterations of annotation to improve the annotation guidelines and, as a result, the developed corpora.

The traditional approach to corpus creation, which is still most frequently used in corpus linguistics, follows a linear process which starts with textual data being collected and an annotation guidelines being created or selected. The data is then annotated and the developed corpus is queried for patterns and features of annotated linguistic categories, e.g. for frequencies of argument supports in a corpus annotated for argumentation.

Agile corpus creation replaces this approach with an iterative small-step process. It starts with a corpus query which allows for the compilation of a small prototypical corpus to set requirements for what should be annotated. This is then followed by the development or selection of an annotation guidelines, and then by the annotation of textual data. Next the results of annotation are evaluated to identify at an early stage design errors, annotation errors and conceptual inadequacies in the guidelines. This is then passed onto the second iteration which goes back to the corpus query phase informed by the analysis of the first iteration. The query can reveal low reliability of annotation (low Inter-Annotator Agreement) which leads to a redesign of the guidelines and re-annotation of textual data and its analysis. The process is repeated, until the analysis demonstrates a satisfactory level of annotation reliability. In what follows, we show how this process of corpus creation can be applied to empirically grounding a theory developed in rhetoric.

4 Bottom-Up Methodology

The paper takes the process of agile corpus creation as the inspiration for proposing a bottom-up methodology for refining a rhetorical theory to the practice of language use which can be further used in linguistic and computational applications. We show the process of such a methodology with the case of the Aristotelian theory of ethos elements. We first describe the overall process (Sect. 4.1); followed by the description of the first iteration of the process which takes a model of Aristotelian ethos elements as an input (Sect. 4.2); its refinements in the second iteration of annotation 4.3; and an empirically-grounded model of ethos elements in the third iteration 4.4.

4.1 Process of Bottom-Up Methodology for Studying Ethos Elements

We propose to employ the process of agile corpus creation to empirically ground a rhetorical theory rather than to develop annotation guidelines, as it has been originally assumed. We modify this process so that it fits the intended purpose (in this section) and show in detail how it can be executed in three iterative steps (the next sections).

The overall process of bottom-up methodology consists of several iterations of sub-processes (in this case use of three iterations; see Fig. 3). First, the initial set of annotation guidelines is built upon the *Aristotelian model* of ethos elements to re-annotate the *EthosHansard2* corpus with the more fine-grained labels for Wisdom, Virtue and Goodwill. This annotation was then evaluated to determine whether

Fig. 3 The process of bottom-up methodology applied to refine a theoretical model to empirically-grounded model

it obtained a satisfactory Inter-Annotator Agreement (IAA). Since it was deemed unsatisfactory—the errors in the corpus are identified and analysed in detail to reveal problems in the annotation guidelines that contributed to the low IAA value.

The second iteration takes the conclusions from the error analysis to *refine* the annotation guidelines which are then reapplied to the same textual data. This means that the refined version of guidelines consists of reinterpreted Aristotelian notions of ethos elements. Since the evaluation of the annotation was still unsatisfactory, a second round of error analysis was performed.

This cycle can be repeated as many times as necessary to obtain satisfactory reliability. In this study, we decided to finish at a third iteration, as the IAA score was satisfactory for such a challenging task. This means that the annotation guidelines of the third iteration contain empirically-grounded notions of ethos elements which can be then explored to identify features and patterns such as, e.g., which ethos element is used by speakers when they support others and which is selected most typically for attacks.

The case study revealed prototypical challenges that can make rhetorical theories difficult to apply in linguistic and computational analysis of the practice of language use. For example, such theories might contain concepts which involve *intentionality* which makes it hard to reliably annotate, e.g., Is a group mentioned in an utterance an audience for the speaker?; How do you reliably decide it, if it is not explicitly stated in the utterance? Another challenge we identified is the *overlap* between concepts in a theory which makes it probable that annotators will decide to select different labels, e.g., Is it always possible to distinguish whether a speaker is supported or attacked for their moral virtues or epistemic/practical virtues? Is it possible to draw a clear line between wisdom and virtue, when wisdom is in fact a type of virtue? In fact, the distinction between wisdom and virtue is not clear even when different

works of Aristotle are considered: in his *Ethics* he excludes intellectual virtues, such as prudence and wisdom, from ethos, while he includes them in his *Rhetoric*.

The three iterations of the process of the bottom-up methodology, applied for the case of ethos elements, are described in detail in the next sections.

4.2 First Iteration of Bottom-Up Methodology with a Theoretical Model of Ethos Elements

Our first attempt at annotating ethos elements has shown that (i) the Aristotelian description of ethos elements is a first step towards designing fine-grained annotation guidelines, but, (ii) it is too generic to capture some distinctive features of ethotic discourse units. The lessons we learned helped us to propose improvements for the second iteration of annotation.

4.2.1 Annotation Guidelines

Building on the original Aristotelian understanding of ethos elements (Aristotle 1991) extended with insights to ethos in Garver (1994), Crowley and Hawhee (2004), Fahnstock and Secor (2003), we proposed the annotation using three main labels (Wisdom, Virtue, and Goodwill) split into support and attack (Argument and Conflict) according to the following guidelines (see <http://arg.tech/f/EthosHansard2WVG1> for full guidelines):

Wisdom Argument From Wisdom (W^+) should be annotated when an entity: (i) is said to have sufficient knowledge for the purpose at hand; or (ii) can draw conclusions from this knowledge; or (iii) has practical experience; or (iv) can draw conclusions from this experience. While Conflict From Wisdom (W^-) should be annotated when the opposite is the case.

Virtue Argument From Virtue (V^+) should be annotated when: (i) a statement refers to the character traits such as, e.g., an entity, when the entity shows positive morality, calmness, justness, selflessness, gracefulness, nobility, positive contributions, liberality, magnanimity or magnificence; or (ii) when an entity provides the correct information. While Conflict From Moral Virtue (V^-) should be annotated when the opposite is the case.

Goodwill Argument From Goodwill (G^+) should be annotated when: (i) a statement refers to an entity's ability to show goodwill to others; or (ii) an entity gives sound advice when it is known, ensuring the entity does not deceive while being inclusive; or (iii) an entity aligns with an audiences values, displaying self sacrifice. While Conflict From Goodwill (G^-) should be annotated when the opposite is the case.

Table 2 Distribution of WVG[±] labels in the *EthosHansard_WVG1* corpus

Labels Count	Supports		Attacks		Total	
	#	%	#	%	#	%
Wisdom (W)	47	28	194	<i>41</i>	241	38
Virtue (V)	104	62	200	<i>43</i>	304	48
Goodwill (G)	18	10	75	16	93	14
Total	169	100	469	100	638	100

Cells highlighted in italic mean the largest proportion of ethos element in ethos appeals (supports or attacks)

4.2.2 Results and Evaluation

The first iteration of annotation according to the guidelines specified in the previous section resulted in the creation of the *EthosHansard2_WVG1* corpus (available at <http://corpora.aifdb.org/EthosHansard2WVG1>) with the distribution of labels such as in Table 2. The most frequently used ethos element in supports was Virtue (63%), while in attacks speakers tend to use Wisdom and Virtue equally frequently (41% and 43%, respectively). Overall according to this annotation, speakers appealed the most commonly to Virtue (38%), and the least frequently—to Goodwill (14%).

A 12% subset of the data was annotated by a second annotator for the purpose of evaluation. Overall this gave Cohen's $\kappa = 0.42$ which, according to (Landis and Koch 1977), is considered fair. As the kappa score is a measurement which sometimes over-penalises for some types of data (Sim and Wright 2005), we decided to also use a simpler measure of percentage agreement (accuracy). Accuracy for our corpus was 57% which we assessed as not reliable enough to annotate ethos elements in conversational argumentation and to treat the results in Table 2 as satisfactorily grounded.

4.2.3 Error Analysis

We identified three types of problems with the annotation guidelines for the first iteration of annotation: (i) when the referent-speaker was supported or attacked due to their actions, the Aristotelian account seemed to limit it to supporting or attacking Wisdom; (ii) when the referent-speaker was accused of lying (and analogically—supported for telling the truth), the initial model limited the choice for annotating it as Virtue; and (iii) when the referent-speaker was accused of not sharing, i.e. withdrawing information (and analogically—supported for sharing), the initial model limited the choice for annotating it to Goodwill. The analysis let us propose solutions for the improved guidelines in the second iteration.

In Example (3), Mr. Lawson is supported by Mrs. Chalker for continued actions (repeated proposals over the longer period) which brought about something good (solved the problems) to a group of people (middle income debtors). This support

was labelled differently by two annotators: one annotator decided that (3) is a support of Wisdom, while another annotator—that it was a support of Goodwill (see labels in the brackets).

- (3) Mrs. Chalker: *The Chancellor of the Exchequer's* [Mr. Lawson's] **proposals** last year, earlier this year, and again at the meeting in Berlin have been fundamental in persuading a change of mind in some other Governments towards **solving the problems of middle income debt-ors** ($W^+ \mid G^+$)

The problem here is that the initial definitions of WVG in the annotation guidelines limited the support for or attack of performing an action just to Wisdom. On the other hand, the actions in (3) brought something good to a group of people which indicates Goodwill rather than Wisdom. We concluded that a person can be supported or attacked for their actions in each case: either when they act wisely (W), or when they act virtuously (V), or they act in goodwill for someone else (G). The scope of the definitions of WVG have then been extended accordingly in the annotation guidelines in the second iteration of the bottom-up methodology.

In Example (4), the phrase 'can you honestly confirm what you have just said' may be interpreted as an accusation of lying, which the initial model treats as an attack on Virtue. Yet the second annotator analysed this expression as an attack on Goodwill. We assumed that the mention of some groups (London and metropolitan countries) might have been interpreted as that the lying mislead these people which could make the annotator label (4) as G^-

- (4) Dr. Cunningham: *Can the Secretary of State* [Mr. Jenkin] **honestly confirm what he has just said** – that there is overwhelming support in **London and the metropolitan countries** for the Government's policy? ($G^- \mid V^-$)

Regardless whether or not this example can be rightly considered as an attack on Goodwill, we decided that it needed to be clarified in the next iteration of guidelines that there can be two cases: (i) lying in general (e.g. "You are the liar") in which case V should be selected; and (ii) lying to others (e.g. "You lied to me") in which case G should be annotated. The similar should be specified for telling truth in the case of support.

In Example (5), the fragment 'will he concede' can be interpreted as an accusation that Mr. Rifkind does not give sound advice when he knows it which would fall under G^- . Yet the second annotator decided to label it as V^- . We hypothesised that the earlier fragment 'Whatever your deeply held views' could suggest the interpretation that by giving this advice, Mr. Rifkind lied, as he does not really believe this advice (his deeply held views are different than advice).

- (5) Mr. Cook: **Whatever** the Minister's [Mr. Rifkind's] **deeply held views** on the matter may be, **will he concede** that the majority of local authorities are appalled

at the prospect of having to sell newly constructed council houses at half price
 ($G^- \mid V^-$)

Thus, we decided that it should be clarified in the guidelines that: (i) lying is to say p , while knowing *not- p* ; and (ii) not sharing is to not say p , while knowing p . Sharing can be also distinguished in terms of ethos elements in the same way as lying: sharing in general (e.g. “You always are willing to give sound advice”) in which case the label V should be selected; and sharing with others (e.g. “You always give me such good advice”) in which case G should be annotated.

4.3 Second Iteration of Bottom-Up Methodology with a Refined Model of Ethos Elements

The output of the sub-process run in the first iteration is then the input for the second iteration which starts with the improvement of annotation guidelines.

4.3.1 Annotation Guidelines

The solutions to the problems identified in the error analysis in Sect. 4.2.3 led to the update of the guidelines for WVG (see <https://www.arg.tech/f/EthosHansard2WVG2> for full guide). The classification was improved by the clear distinction between: (i) information p vs action a ; (ii) lying as knowing p , but saying *not- p* vs not sharing as knowing p , but not saying p ; and (iii) saying/doing in general in the case of Virtue vs saying/doing to an audience in the case of Goodwill. As a result, the labels WVG are defined as follows:

Wisdom Argument From Wisdom should be annotated when an entity: (i) knows the right information p ; (ii) knows the right action a . Conflict From Wisdom should be annotated when an entity: (i) does not know p ; or (ii) does not do a .

Virtue Argument From Virtue should be annotated when an entity: (i) knows and reveals the right information in general (knows p and says p); or (ii) does not lie in general (knows p and does not say *not- p*); or (iii) does the right thing in general (knows a and does a); or (iv) does not do the wrong action in general (knows a and does not do *not- a*). Conflict From Virtue should be annotated when an entity: (i) knows information but does not reveal it in general (knows p and does not say p); or (ii) lies in general (knows p and says *not- p*); or (iii) does not do the right thing in general (knows a and does *not- a*); or (iv) does the wrong action in general (knows a and does *not- a*).

Goodwill Argument From Goodwill should be annotated when an entity: (i) knows and shares information with the audience (knows p and says p); or (ii) doesn't mislead the audience (knows p and does not say *not- p*); or (iii) performs the right action for others aligning with their values giving sound advice (knows a and does a); or (iv) does not do wrong to others (knows a and does not do *not- a*). Conflict From Goodwill should be annotated when an entity: (i) does not share information with the audience (knows p and does not say p); or (ii) misleads the audience (knows p and says *not- p*); or (iii) does not do what they know is right for the audience

Table 3 Distribution of the WVG+/- labels in the *EthosHansard2_WVG2* corpus

Labels Count	ESE+		ESE-		Total	
	#	%	#	%	#	%
Wisdom (W)	49	29	189	<i>40</i>	238	37
Virtue (V)	100	59	193	<i>41</i>	293	<i>46</i>
Goodwill (G)	20	12	87	19	107	17
Total	169	100	469	100	638	100

Cells highlighted in *italics* mean the largest proportion of ethos element in ethos appeals

(knows *a* and does not do *a*); or (iv) does the wrong things for the audience (knows *a* and does *not-a*).

4.3.2 Results and Evaluation

The second iteration of annotation according to improved guidelines resulted in the creation of the *EthosHansard2_WVG2* corpus (available at <http://corpora.aifdb.org/EthosHansard2WVG2>) with the distribution of labels in Table 3. The tendencies in this iteration are very similar to the tendencies in the previous iteration. First, the most frequently used ethos element in supports is still Virtue (59% as opposed to 63% in the first iteration), while in attacks speakers again tend to use Wisdom and Virtue equally frequently (40% and 41% as opposed to 41% and 43% in the first iteration). Overall, speakers appealed most commonly to Virtue (46% vs 38% in the first iteration), and least frequently—to Goodwill (17% vs 14% in the first iteration).

To evaluate the improved guidelines, a 12% subset of the corpus was again annotated by the second annotator giving $\kappa = 0.52$ and a percentage agreement of 66% as opposed to $\kappa = 0.42$ and accuracy of 57% in the first iteration. This means a 24% improvement in terms of Cohen's kappa and a 16% improvement in terms of accuracy. Still we decided that the reliability of annotation was not satisfactory yet, and we should iterate at least one more time.

4.3.3 Error Analysis

In this iteration of annotation, we grouped errors into five types of problems: (i) intentionality of a concept of an audience; (ii) ambiguity of words related to the WVG labels; (iii) overlap between the WVG labels; (iv) the lack of external knowledge; and (v) granularity of segmentation.

In Example (6), Mr. Jenkin attacks the Prime Minister as not acting in the best interest of specific groups (represented by European leaders) and as not aligned with anyone ('is a solitary one'). Such an interpretation would fall under label of G^- , yet the second annotator decided to select V^- .

- (6) Mr. Jenkin said, *Is it not the case that, despite what the hon. Member for Eastbourne [Mr. Gow] said earlier, the Prime Minister's Bruges approach is a **solidary one** and is **not supported by any other European leader*** ($G^- \mid V^-$)

We hypothesised that the problem lies in the fact that Goodwill is associated in the guidelines with an audience and it might be difficult here to decide whether or not these groups are in fact the Prime Minister's *intended audience*, or they are some groups who happened to be brought about by Mr. Jenkin. If the attack was formulated as "the Prime Minister's Bruges approach is bad for the EU nations including the British people", it would be quite clearly indicated that the Prime Minister is considered to not act in the best interest of his audience, as the British people should be his targeted audience. But this type of clear indication might be rarely surfaced or even omitted on purpose. We decided that we should avoid to use "audience" in the guidelines at all, and instead use "others" which means any group that are an intended audience or just happened to be mentioned.

Example (7) illustrates a problem of using some words or phrases in an *ambiguous* or *uncommon* way. The use of the word 'expert' here might have been a shortcut for one of the annotators to select Wisdom, while another annotator noticed the untypical meaning of this word used ("expert in doggerel and verse") and selected Virtue.

- (7) Dr. Cunningham: *I understand the right hon. Gentleman [Mr. Baker] to be something of an **expert in doggerel and verse*** ($V^- \mid W^-$)

Typically, such keywords are treated in guidelines as a tool allowing for a quick identification of labels. Yet, they can be used in an idiomatic, ironic or interrogative manner which will deviate from their standard meaning. To solve this problem, we proposed to add highlighted notes in the guidelines (boxed with a italic background) which was then used anytime we needed to draw annotators' attention to some exceptions or peculiarities.

The rest of the problems were solved in the same manner by introducing polymorphic labels. Lets consider these errors first. Example (8) shows the problem that ethos elements *overlap* in some instances, i.e. there is no clear borderline between WVG. The decision which of the two labels (W^- or V^-) should be annotated depends on the interpretation of what 'living in never-never land' means: whether Mr. Chalker refers to Mr. Aitken's deliberate action ('living in never-never land' interpreted as deliberately misleading people, i.e. V^-) or to acting not on purpose (living in the world of fantasy which is not deliberate, i.e. W^-).

- (8) Mr. Chalker: *I will really think that he [Mr. Aitken] is **living in never-never land*** ($W^- \mid V^-$)

Example (9) demonstrates a problem of *external knowledge* not being available for annotators to properly reconstruct the meaning of the utterance and select a specific label. Here, Mr. Skinner is accused of an inability to say something about

his leader. Yet depending on whether this inability is caused by withholding the information or by a lack of knowledge of his leader, this expression should be annotated as V^- or W^+ , respectively.

- (9) Mr. Prior: *I suspect that **that is more than** the hon. Gentleman [Mr. Skinner] **can say about his leader**.* ($V^- \mid W^+$)

Example (10) illustrates the problem of the *granularity of segmentation*. In this annotation, the segmentation on the sentence level rather than the clause level was selected. Sentences are relatively easy to detect not only manually, but also automatically. An alternative such as EDUs (Elementary Discourse Units) or ADUs (Argumentative Discourse Units) are hard to recognise, as they require the identification of functional or intentional relations between fragments of texts. On the other hand, a sentence can be complex and carry two or more types of attacks or supports. (10) is composed of two supports of the Government for doing the right thing (W^+) and for showing the virtue of Europeanness (V^+). It just happened that one annotator picked up on one support and the second annotator focused on another support, but in fact both of these supports are present in this expression.

- (10) Mr. Taylor said, *Has my right hon. Friend taken note of the various measures that the Government **have already introduced**, which are well ahead of those of our European friends in terms of **showing our Europeanness*** ($W^+ \mid V^+$)

No matter how much we could improve the guidelines and how much better annotators were trained, the last three problems cannot be easily resolved. In the case of the overlap, if an ethotic expression falls into a italic area of the overlap between, e.g., W and V , then it should not be arbitrarily decided that it is, e.g. W but not V . For the lack of external knowledge, if the meaning of an ethotic expression is dependent on a reconstruction necessary to decide whether it is, e.g., W or G , and this reconstruction is not available to us, then we should not arbitrarily guess that this is, e.g., G but not W . In the case of granularity of the segmentation, if the segment of an ethotic expression is large (as a trade-off for being easily detectable) and carries both, e.g., V and G , then we should not arbitrarily choose one over another. We propose to solve these problems by introducing four polymorphic labels which contain more than one ethos element, i.e. WV , WG , VG , WVG . As a result, if a decision between a single ethos element is unjustified due to the overlap, external knowledge or segmentation, then the annotator may select such a polymorphic label.

4.4 Third Iteration of Bottom-Up Methodology with an Empirically-Grounded Model of Ethos Elements

The output of the sub-process run in the second iteration is then the input for the third iteration which starts with the improvement of annotation guidelines.

4.4.1 Annotation Guidelines

The addition of the polymorphic types made the guidelines even more complex, as the more labels an annotator can choose from the more cognitively difficult the task becomes. Apart from specific refinements such as adding highlighted notes with the explanations of exceptions or peculiarities, the main solution for the new guidelines included: (i) replacing the linear description of labels with the description through decision trees with questions to guide annotators to the right decisions; (ii) introducing polymorphic types; (iii) illustrating and explaining many descriptions with examples. The resulting guidelines were detailed with a total of 21 rules and several sub-rules.

For ethos supports, two decision trees were introduced (for non-polymorphic labels and for polymorphic labels) which start with a question about a label that is the easiest to annotate, followed by other labels in a process of elimination. For non-polymorphic types, it starts with G as it is relatively easy to determine whether an ethotic expression refers to a group of people, i.e. to “others”. The decision tree starts with the question: **(Rule 3.1.1)**¹ *Does the source-speaker say that the referent-speaker aligns with others?* If the answer to this question is “Yes”, the annotator should choose G⁺. If the answer is “No”, she moves to the next question: **(Rule 3.1.2)** *Does the source-speaker say that the referent-speaker says the truth to others or she doesn't lie to others?* If G is excluded, the next question covers W, and finally - V (there are seven rules in this decision tree in total), on the assumption that V is the largest and most vague category which should be selected anyway, if G⁺ and W⁺ are eliminated.

The rules are then further specified and exemplified. For example, Rule 3.1.1 has the following description: “This question makes you consider whether the ethotic expression states that the source-speaker X supports the referent-speaker Y, because **Y aligns herself with some group of people** by stressing that she is ‘one of them’. For example, a politician Y might align herself with citizens, voters, local community, minority, etc by presenting herself as one of them, as equal to them, as close to their needs, as understanding their needs etc. In such cases, you should annotate this ethotic expression as a support on the grounds of goodwill, i.e. G⁺.”

The polymorphic labels are introduced through the decision tree as well (with the total of four rules). The formulation of questions is straightforward, as they ask about components of a polymorphic label. For example, the first rule considers WG⁺ and asks the annotator to decide: **(Rule 3.1.8)** *Does the ethos support fall under W⁺ or G⁺?* The next polymorphic labels to decide are VG⁺, WV⁺, and finally - WVG⁺.

¹ Number 3 at the beginning of the rule means the third iteration, and number 1 on the second place—the rule for supports.

Table 4 Distribution of the WVG+/- labels in the *EthosHansard2_WVG3* corpus

Labels Count	Supports		Attacks		Total	
	#	%	#	%	#	%
Wisdom (W)	109	<i>64.5</i>	128	27.3	237	<i>37.1</i>
Virtue (V)	24	14.2	163	<i>34.8</i>	187	29.3
Goodwill (G)	20	11.8	73	15.6	94	14.7
Non-polymorphic	153	90.5	364	77.6	517	81
WV	13	<i>7.7</i>	78	<i>16.6</i>	91	<i>14.3</i>
WG	1	0.6	7	1.5	8	1.3
VG	–	–	11	2.3	11	1.7
WVG	2	1.2	9	1.9	11	1.7
Polymorphic	16	9.5	105	22.4	121	19
Default	–	–	–	–	–	–
Total	169	100	469	100	638	100

Cells highlighted in italic mean the largest proportion of ethos element in ethos appeals (separately for non-polymorphic and polymorphic types)

Table 5 Five interpretations of inter-annotator agreement in the *EthosHansard2_WVG3* corpus which address the issue of over-penalising disagreement between annotators resulting from polymorphic types

Inter-annotator agreement	Accuracy	Cohen's κ
Non-polymorphic measure	62	.49
Non-overlapping measure	64	.55
Conservative overlapping measure	54	.43
Moderate overlapping measure	62	.48
Liberal overlapping measure	70	.54

The rules for ethos attacks are created in the similar manner. See <http://arg.tech/f/EthosHansard2WVG3> for full guidelines.

4.4.2 Results and Evaluation

The third iteration of annotation according to the final set of guidelines resulted in the *EthosHansard2_WVG3* corpus (available at <http://corpora.aifdb.org/EthosHansard2WVG3>) with the distribution of labels such as in Table 4. The majority of ethotic expressions belong to the non-polymorphic types (81%), while polymorphic types constitute 19% of all annotated labels. This indicates that their inclusion has turned out to be crucial for obtaining an adequate picture of the relations between Wisdom, Virtue and Goodwill.

Table 6 The comparison of inter-annotator agreement at each iteration of annotation of ethos elements in Hansard data

Annotation iteration	Accuracy	Cohen's κ
WVG1	57	.42
WVG2	65	.52
WVG3 non-overlapping	64	.55
WVG3 polymorphic	70	.54

For *EthosHansard2_WVG3* corpus, we give the IAA interpretation which over-penalises for the polymorphic types to the least extent. Cells highlighted in italic mean the highest IAA

The tendencies in this iteration turned out to alter from the previous iterations. The most frequently used ethos element in supports is now observed to be W (64.5%) rather than V (63% in the first iteration and 59% in the second). In attacks, V became noticeably more frequent than W (34.8% vs 27.3%) as opposed to the previous iterations which indicated that they are almost equally frequent (41% and 43% in the first iteration and 40% and 41% in the second). Overall, speakers appealed the most commonly to W rather than to V as observed in the previous iterations, but still the least frequently to G (14.7% here, and 14% in the first iteration and 17% in the second). For polymorphic types, the dominant tendency is observed for the appeals to WV with the even distribution between the rest of the three types.

To evaluate the improved guidelines, a 12% subset of the corpus was again annotated by the second annotator. In this case, the evaluation has not been straightforward though, as decisions made with respect to 14 labels in total (3 non-polymorphic and 4 polymorphic multiplied by 2 for supports and attacks) is much harder than decisions made with respect to 6 labels. The task is even harder in where the polymorphic labels needed to be introduced, as 14 completely different labels are still easier to annotate than 14 labels which overlap. For example, if one annotator will select WV^+ and another selects VG^+ , then should it be considered as full disagreement or partial agreement? Thus, we applied various measures of Inter-Annotator Agreement (IAA) to have a range of ways to compare the reliability of this iteration with the previous ones, trying to reduce over-penalisation for using polymorphic types as much as possible (see Table 5).

The *non-polymorphic measure* excludes any instances of a label which has been annotated as polymorphic type. The advantage of this solution is that it allows for the direct comparison of this iteration of annotation to the previous ones. Yet it results in the significant reduction of the size of the sample of corpus—from 12 to 9.4% which means that one disagreement will penalise IAA to larger extent.² Thus, we decided to look at other ways of evaluating this iteration.

The *non-overlapping measure* excludes instances of disagreement, when ethos elements were overlapping (e.g. for W and WV, or WV and VG, but leaves ethotic expressions, when there is an agreement on a polymorphic type, e.g. WV^- and WV^-). As a result, we reduced the sample of IAA annotation to 11% rather than to 9.4% in comparison to 12% of the sample in the previous iteration. This means that

² If two annotators disagree once in a 2 element corpus, then the accuracy is 50%, while in the corpus of size 10, the accuracy will be 80%.

it is a better trade-off between comparability of iterations (we reduce the uncertainty of annotation coming from the overlap) and the size of samples of annotation to compare. In this case, $\kappa = 0.55$ and accuracy of 64% gives an improvement over the previous iterations (Table 6).

The next three measures fully account for polymorphic types. The most harsh measure is the *conservative interpretation* which assigns 0 to any disagreement, i.e. the IAA annotation of an ethotic expression will score 0, if two annotators will choose W^+ and V^+ , but also if they chose WV^+ and VG^+ (or even WV^+ and WVG^+). *Moderate interpretation* assigns 0.5 to overlaps, while *liberal interpretation* assigns them 1. This means that in the last case, only full disagreement will be penalised (e.g. W^+ and V^+ or WV^+ and G^+ will score 0), which is similar to the non-polymorphic case in this sense that 0 is supposed to be assigned to categories which have nothing in common (e.g. to W^+ and V^+). Table 6 shows that the liberal measure gives an improvement over the previous iterations both in terms of accuracy and kappa.

5 Discussion

Section 4 introduced a novel bottom-up methodology which allows for empirically grounding a theoretical model in the practice of language use. The process was demonstrated on the selected case of appeals to ethos elements, yet it is possible to formulate suggestions which provide further guidance of how to execute this method to other theories of conversational argumentation. What makes it possible is that the problems identified in the error analyses are not ethos dependant, but rather depend on the issues related to how theories are built, and the issues related to linguistic features of language use.

In terms of issues coming from theories, our case study identified three problems: (i) non-exhaustive specification of terms in a theory; (ii) overlapping terms; (iii) intentionality of concepts used to specify terms. In our study, the first problem has surfaced for the specification of ethos elements with respect to actions. Possibly because actions manifest themselves mostly through experience—a component of practical wisdom in Aristotelian rhetoric—they have not been mentioned in the context of the other two ethos elements. Though, in as much as W can be specified as “knowing p ” (having practical knowledge) or “knowing a ” (having practical experience), it should be allowed for V (as well for G) to be specified both in terms of information and action, i.e. as “knowing p and not saying *not-p*” (not lying) or “knowing a and not doing *not-a*” (not doing wrong action).

Another challenge which we might encounter in applying this method can be that (at least) some terms in a theory overlap. A similar problem to ethos elements is often mentioned with respect to the even more central distinction in rhetoric of logos-ethos-pathos (cf. Hinton and Budzyńska-Daca 2019). For example, in ad verecundiam argument a speaker aims to invoke an emotion of shame in the audience/

opponent (pathos) by appealing to an authority (ethos) who claims something contrary to what the audience/opponent claims.

Finally, the specification of some terms can include intentional concepts such as audience or belief. Formal dialogue systems, for example, introduced the concept of ‘commitment’, which means ‘publicly declared belief’, to replace ‘belief’. As a result, instead of using a concept which requires to have an insight into the internal state of an agent, such systems use a concept which requires to know only what has been said in a dialogue in order to determine who is committed to what. In a similar way, in order to determine whether a group is an intended audience for a given speaker would require to know an internal state of this speaker. Such a concept might not be a problem for a theory, but becomes one in the practical applications of this theory, if an annotator needs to decide whether a group is an audience for a given speaker, and this is not expressed explicitly on the linguistic surface of the ethotic expression.

In terms of linguistic issues, our case study identified three problems: (i) ambiguity of words used in ethotic expressions; (ii) the lack of external knowledge required to reconstruct the meaning of an ethotic expression; (iii) granularity of segmentation of ethotic expressions into units. In our case study, the first problem was encountered in one of the ethotic expressions in the corpus in which the word ‘expert’ was used in an unusual way. In both manual and automatic annotation, it is very common to use keywords (or bag of words) or discourse markers to enhance the identification of specific categories. In our case, ‘expert’ is a natural candidate to be a keyword which signals that the speaker means W. Yet, unusual (idiomatic, ironic or interrogative) use will change the standard meaning of a keyword and the guidelines should somehow account for it.

Another challenge is the lack of external knowledge, if the meaning of an ethotic expression requires reconstruction. In natural language this is usually the case, but the degree of how much reconstruction is required will vary. The reconstruction may be not necessary if a coherent and understandable meaning can be identified, e.g. in “The Government has shown Europeanness” we might not be able to reconstruct ‘when’ and ‘where’ it has been shown, but we can understand ‘who’ and ‘what’ has been shown. On the other hand, in “This is more that what he can say about his leader” we can understand that he can’t say more about his leader, but we are unable to identify grounds for this inability (no knowledge about the leader vs hiding the information about the leader) which does not allow an annotator to decide which ethos elements to choose as a label (W vs V, respectively).

Finally, the analysis of the practical use of language requires working with a specific building block of a text. That is, it requires a decision to be made on how a text should be segmented, typically on a clause- or sentence-level. The second type of building block is easier technically, because it is relatively easy to specify boundaries of a sentence (starting with a capital letter and ending with a full stop). Although in this case, one sentence can contain more than one category which has been analysed, such as in our example when in one sentence a speaker was supporting W and V of the Government.

6 Conclusions

In the paper we have shown how to do empirical studies informed by theory. Having observed the problem of the direct applicability of theories to study the practice of language use for linguistic and computational applications, we proposed a novel bottom-up methodology that helps to refine existing conceptual frameworks to the study of natural language data. In other words, we have shown the method of adapting existing theories to capture the richness and complexity of natural argumentation. For that purpose we studied ethotic appeals in political debates by taking the Aristotelian account of ethos elements: practical wisdom, moral virtue and goodwill, and we have shown how to apply the process of agile corpus creation to study conversational argumentation. Through the iterative annotation and evaluation process, we developed an empirically-grounded model of appeals to ethos elements which can then be used to gain insight into patterns in the use of these appeals in real practice, such as that politicians tend to use appeals to Wisdom when they support each other, but choose rather Virtue for an ethotic attack. This study shows that the efforts of making an application of theories to data realistic requires a permanent ‘dialogue’ between conceptual frameworks and the empirical insights.

Acknowledgements The work reported in this paper was supported in part by the Polish National Science Centre under Grant 2015/18/M/HS1/00620 and in part by POB CyberDS of Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme under Grant 1820/1/Z01/POB3/2021.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Araszkiewicz, M., and M. Koszowy, 2016. Deontic authority in legal argumentation: A case study. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, volume 1, pp. 1–19. London: College Publications.
- Aristotle. 1991. *On Rhetoric* (G. A. Kennedy, Trans.). New York: Oxford University Press.
- Barth, E.M., and E.C.W. Krabbe. 1982. *From Axiom to Dialogue: A Philosophical Study of Logics and Argumentation*. Berlin: De Gruyter.
- Baumlin, J.S., and P.L. Scisco, 2018. Ethos and its Constitutive Role in Organizational Rhetoric. In *The Handbook of Organizational Rhetoric and Communication*, ed. Oyvind Ihlen and Robert L. Heath, pp. 201–2014. Wiley.
- Brinton, A. 1986. Ethotic Argument. *History of Philosophy Quarterly* 3: 245–257.
- Budzynska, K. (2010). Argument Analysis: Components of Interpersonal Argumentation. In *Frontiers in Artificial Intelligence and Applications. Proceedings of 3rd International Conference on Computational Models of Argument (COMMA 2010)*, pp. 135–146.
- Budzynska, K., and Reed, C. 2011. Whence inference. Technical report, University of Dundee.

- Budzynska, K., M. Janier, C. Reed, and P. Saint-Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument and Computation* 7(1): 91–108.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.
- Crowley, S., and D. Hawhee. 2004. *Ancient Rhetorics for Contemporary Students*. London: Pearson/Longman.
- Duthie, R. and K. Budzynska. 2018a. A deep modular RNN approach for ethos mining. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 4041–4047. International Joint Conferences on Artificial Intelligence Organization, vol. 7. <https://doi.org/10.24963/ijcai.2018/562>.
- Duthie, R. and K. Budzynska 2018b. Classifying types of ethos supports and attacks. In *Proceedings of the 7th International Conference on Computational Models of Argument (COMMA 2018)*, pp. 161–168.
- El Baff, R., H. Wachsmuth, K. Al Khatib, M. Stede, and B. Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on NLG*, pp. 54–64.
- Fahnestock, J., and M. Secor. 2003. *A Rhetoric of Argument*. New York: McGraw-Hill Higher Education.
- Garver, E. 1994. *Aristotle's Rhetoric: An Art of Character. Philosophy/Classics/Rhetoric*. Chicago: University of Chicago Press.
- Goldman, A.I. 2001. Experts: Which ones should you trust? *Philosophy and Phenomenological Research* 63(1): 85–110.
- Goodwin, J. 2011. Accounting for the appeal to the authority of experts. *Argumentation* 25: 285–296.
- Goodwin, J., and V. Cortes. 2010. Theorists' and practitioners' spatial metaphors for argumentation: A corpus-based approach. *Verbum* 22: 163–178.
- Groarke, L., and C. Tindale. 2013. *Good Reasoning Matters! A Constructive Approach to Critical Thinking*. Oxford: Oxford University Press.
- Habernal, I., H. Wachsmuth, I. Gurevych, and B. Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in Web argumentation. In *Proceedings of the Conference of the NAACL*, pp. 386–396.
- Hansen, H.W. 2019. *The Stanford Encyclopedia of Philosophy, Chapter Fallacies*. New York: Wiley.
- Hardwig, J. 1991. The role of trust in knowledge. *Journal of Philosophy* 88(12): 693–708.
- Hidey, C., E. Musi, A. Hwang, S. Muresan, and K. McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pp. 11–21.
- Hinton, M. 2021. Corpus linguistics methods in the study of (meta)argumentation. *Argumentation* 35: 435–455.
- Hinton, M., and A. Budzyńska-Daca. 2019. A comparative study of political communication in televised pre-election debates in Poland and the United States of America. *Research in Language* 17(1): 1–19.
- Jackson, S. 2015. Deference, distrust, and delegation: Three design hypotheses. In *Reflections on Theoretical Issues in Argumentation Theory*, ed. F.H. van Eemeren and B. Garssen, 227–243. Berlin: Springer.
- Janier, M., J. Lawrence, and C. Reed. 2014. OVA+: An argument analysis interface. In *Computational Models of Argument (COMMA)*, pp. 463–464.
- Koszowy, M., and D. Walton. 2017. Profiles of dialogue for repairing faults in arguments from expert opinion. *Logic and Logical Philosophy* 26(1): 79–113.
- Koszowy, M., and D. Walton. 2019. Epistemic and deontic authority in the argumentum ad verecundiam. *Pragmatics and Society* 10(2): 151–179.
- Landis, J.R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.
- Lawrence, J., and C. Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45(4): 765–818.
- Lawrence, J., M. Janier and C. Reed. 2015. Working with open argument corpora. In *Proceedings of the European Conference on Argumentation (ECA)*, volume 1, pp. 367–380.
- Liao, Y. 2021. The legitimacy crisis of arguments from expert opinion: Can't we trust experts? *Argumentation* 35(2): 265–286.
- Lorenz, K., and P. Lorenzen. 1978. *Dialogische Logik*. Darmstadt: WBG.
- Macagno, F. 2013. Strategies of character attack. *Argumentation* 27(4): 369–401.
- Mizrahi, M. 2013. Why arguments from expert opinion are weak arguments. *Informal Logic* 33(1): 57–79.

- Musi, E., and M. Aakhus. 2019. Framing fracking: Semantic frames as meta-argumentative indicators for knowledge-driven argument mining of controversies. *Journal of Argumentation in Context* 8(1): 112–135.
- Musi, E., D. Ghosh, and S. Muresan. 2016. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the third workshop on argument mining (ArgMining2016)*, pp. 82–93.
- Pereira-Fariña, M., M. Koszowy, and K. Budzynska. 2022. 'It was never just about the statue': Ethos of historical figures in public debates on contested cultural objects. *Discourse and Society* 33(2).
- Rapp, C. 2010. Aristotle's rhetoric. *The Stanford Encyclopedia of Philosophy*.
- Seidel, M. 2014. Throwing the baby out with the water: From reasonably scrutinizing authorities to rampant scepticism about expertise. *Informal Logic* 34(2): 192–218.
- Sim, J., and C.C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85(3): 257–268.
- Stede, M., and J. Schneider. 2018. *Argumentation Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, New York
- Tindale, C.W. 2011. Character and knowledge: Learning from the speech of experts. *Argumentation* 25: 341.
- Visser, J., J. Lawrence, C. Reed, J.H.M. Wagemans, and D. Walton. 2021. Annotating argument schemes. *Argumentation* 35(1): 101–139.
- Voormann, H., and U. Gut. 2008. Agile corpus creation. *Corpus Linguistics and Linguistic Theory* 4(2): 235–251.
- Wacholder, N., S. Muresan, D. Ghosh, and M. Aakhus. 2014. Annotating multiparty discourse: Challenges for agreement metrics. In *Proceedings of the 8th Linguistic Annotation Workshop: COLING*, pp. 120–128
- Wachsmuth, H., M. Stede, R. El Baff, K. Al Khatib, M. Skeppstedt, and B. Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3753–3765.
- Wagemans, J.H.M. 2011. The assessment of argumentation from expert opinion. *Argumentation* 25(3): 329–339.
- Walton, D. 1997. *Appeal to Expert Opinion: Arguments from Authority*. University Park: Penn State University Press.
- Walton, D. 1999. Ethotic arguments and fallacies: The credibility function in multi-agent dialogue systems. *Pragmatics & Cognition* 7(1): 177–203.
- Walton, D., and E.C.W. Krabbe. 1995. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. Albany: SUNY Press.
- Walton, D., C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge: Cambridge University Press.
- Wrisley, G. 2019. *Bad Arguments: 100 of the Most Important Fallacies in Western Philosophy*, chapter Ad Hominem: Circumstantial, pp. 77–82. Wiley.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.