4 Intercoder reliability in qualitative discourse analysis

Coline Rondiat, Christoph Niessen, Barbara De Cock, and Philippe Hambye

1 Introduction: reliability and reproducibility in qualitative research

Reliability, in its broadest sense, refers to the consistency and dependability of a measurement or classification process (Carmines & Zeller, 1979). As such, it constitutes a cornerstone of scientific inquiry: without reliable procedures, empirical findings cannot be meaningfully replicated or compared across studies. In recent years, debates about reliability have become more prominent within interpretive research, as scholars seek ways to demonstrate rigor without undermining the contextual and situated nature of qualitative inquiry. Central to this effort is the concern that qualitative interpretations should not be dismissed as merely idiosyncratic or contingent upon individual subjectivity. This issue comes to the fore in qualitative coding, where the demand for methodological rigor has led scholars to distinguish between different forms of reliability. These include: stability, or the extent to which a single coder applies codes consistently over time; accuracy, which assesses the alignment of a coding scheme with a recognised 'gold standard'; and reproducibility, which evaluates the degree of agreement among multiple coders using the same coding protocol (Campbell *et al.*, 2013: 295).

Importantly, reproducibility is deemed to act as a partial safeguard against interpretive idiosyncrasy, insofar as it evidences that independent coders can reach comparable judgments. In practice, reproducibility is most often assessed through intercoder reliability (ICR), a statistical indicator that measures the consistency of coding across different individuals. This assessment typically employs coefficients such as Cohen's κ, Krippendorff's α, or Scott's π, each based on distinct statistical assumptions which may yield marginally divergent outcomes. These coefficients are used to quantify the degree of agreement among two or more coders when classifying the same dataset into predefined categories (Cofie *et al.*, 2022: 73; O'Connor & Joffe, 2020: 2; Wilson-Lopez *et al.*, 2019: 8). When

DOI: 10.4324/9781003610489-5
This chapter has been made available under a CC BY-NC-ND 4.0 license

exceeding an established threshold, ICR is typically taken to indicate that the coding procedure is being implemented in a stable and replicable manner throughout the dataset. For this reason, it has come to function as a widely adopted quality metric across a range of disciplines, often taken as a warrant for the overall procedural reliability of a given stage in the research process (Hruschka *et al*., 2004: 309).

While a growing number of scholars advocate for the explicit assessment of coding reliability – including through ICR – in qualitative research (Campbell *et al*., 2013: 295), critics argue that this development reflects an epistemic importation of positivist assumptions, most notably that of a single, objectively recoverable reality, into interpretivist frameworks that understand meaning as contingent, situated, and co-constructed (Braun & Clarke, 2013; Hruschka *et al*., 2004: 308; Stenbacka, 2001). Such tension often surfaces when coders working within a constructivist paradigm reach divergent yet equally defensible interpretations of an item. In these cases, coder disagreement does not signal a shortcoming in the analytic procedure, but rather reflects the interpretive richness of the data. From this vantage point, the recourse to ICR can be seen as reducing a dialogical and iterative process to a procedural exercise in alignment. Critics therefore identify an 'epistemological problem' (Braun & Clarke, 2013): namely, the conflation of consistency with validity, and the tacit flattening of interpretive complexity to satisfy externally imposed standards of rigor. The notable lack of consensus on how ICR should be operationalised in qualitative contexts (O'Connor & Joffe, 2020: 1) may, in part, reflect this underlying philosophical resistance.

An alternative reading – one increasingly prevalent in multidisciplinary and team-based research contexts – casts ICR not as a positivist imposition but as a pragmatic tool. In this view, ICR is mobilised less as a claim to universal truth than as a device to (i) enhance transparency of coding decisions, (ii) mitigate the influence of individual cognitive or cultural bias, and (iii) facilitate the replicability of interpretive judgments across research teams (Hruschka *et al*., 2004: 309; Krippendorff, 1980; O'Connor & Joffe, 2020: 4). What is sought, then, is not a potentially homogenous or unique interpretation, but the disciplined management of variability in interpretive labor. This pragmatic orientation tends to surface in team-based coding, where ICR is deployed as a means of coordinating standards and adjusting analytic instruments. Divergent readings among coders are not uncommon, and, in response to those, teams usually produce a shared codebook to stabilise interpretations. High ICR is hence taken as evidence that the codebook provides a solid framework, while low scores are read as signals for refinement. In practice, this usually entails an initial coding round, followed by structured discussions to address discrepancies; the codebook

is then revised and successive rounds are conducted until an acceptable level of agreement is reached (Hruschka *et al.*, 2004: 313). In this way, ICR functions not only as a measure of agreement but also as a diagnostic instrument for improving the clarity and usability of the coding scheme (O'Connor & Joffe, 2020: 3).

Some scholars therefore argue that low agreement should be read primarily as an occasion for reflexivity: disagreement can expose ambiguities or unexamined assumptions within the coding team, thereby prompting deeper reflection on the coding scheme and procedure (O'Connor & Joffe, 2020: 6). When persistent coding issues emerge, several remedial strategies may be implemented. These include eliminating or merging poorly performing codes, revising definitions and coding instructions, or supplementing the codebook with concrete examples (Campbell *et al.*, 2013: 300). There is, however, little consensus on the most appropriate course of action. Ultimately, the choice depends on the researcher's epistemological stance, the aims of the project, and whether ICR is being used as a validation tool or a developmental aid. In highly standardised studies, discarding unreliable codes may be preferred since the priority lies in maintaining measurement reliability and comparability across coders. In contrast, in exploratory or constructivist research, refining and re-testing such codes may better serve the study's interpretive goals, as apparent inconsistencies can reveal meaningful ambiguities in discourse and contribute to the reflexive refinement of analytical categories (O'Connor & Joffe, 2020: 9). What remains consistent across traditions, however, is the importance of making these processes explicit. Best practice in reporting recommends that researchers not only present their ICR statistics but also describe the process undertaken to achieve them, including coding rounds, codebook changes, and decision rationales (Miles & Huberman, 1994). In doing so, ICR is situated within a broader architecture of reliability and validity checks, and functions not as a standalone guarantor of analytic quality but as one element in a transparent, well-documented research process.

Against this backdrop, this chapter examines how ICR can be meaningfully applied within an interpretive, discourse-analytic framework. Using the ICR conducted on the coding of 'populism' and its derivatives (henceforth *populis**) as an illustrative case, it first considers the specific methodological difficulties involved in analysing these terms. It then outlines how the coding grid has been designed to address such issues and describes the procedures used to assess reliability across multiple category types. Finally, it offers an interpretation of the resulting patterns of agreement and divergence. Taken together, these discussions aim to clarify what reliability can reasonably signify within and bring to a discourse-analytic approach.

2 Specific challenges of the study of sociopolitical keywords

The qualitative coding of sociopolitical keywords (henceforth SPKW; see Hambye *et al.*, Chapter 3 in this volume), such as *populis**, entails distinctive methodological difficulties, particularly when the aim is to capture both their explicit and implicit dimensions within public discourse. These challenges bear directly on ICR and, by extension, on the epistemic claims that can be made about the findings.

A first issue concerns data segmentation, which requires balancing two competing demands: preserving the contextual richness in which an SPKW appears, and managing the interpretive complexity entailed by larger coding units. Prior research indicates that coding extended units – full sentences or discourse segments – enhances interpretive validity by retaining contextual cues essential to meaning-making (O'Connor & Joffe, 2020: 6), especially when working with linguistically and semantically dense material. Yet broader segments frequently contain multiple, and at times overlapping, meanings, which increases the potential for interpretive divergence among coders (Hruschka *et al.*, 2004: 319; O'Connor & Joffe, 2020: 6). The study illustrated here focuses on a single lexical item – 'populism' and its morphological variants. All occurrences were interpreted using the entire discourse segment in which they appear, as the analytical priority was to capture their discursive functions and contextual meanings (Hambye *et al.*, Chapter 3 in this volume). Such an approach is indispensable for a semantically and politically informed analysis (Hambye & Rondiat, 2024). When focusing on ICR, however, such extended and heterogeneous units inevitably broaden the scope for interpretive variation (Hruschka *et al.*, 2004).

Interpretive depth constitutes a further challenge. Coders were instructed to give prevalence to close textual engagement and to avoid speculative or excessively inferential readings. Yet, the very task of reconstructing SPKW meaning, including the more implicit nuances conveyed through lexical cues, necessarily demands a fine-grained judgment. As a result, the coding process was concerned not only with identifying surface features, such as word class, but also with reconstructing implicit meaning. For instance, in excerpt (1), 'populism' functions as a definite noun, implying that the phenomenon is already known and readily identifiable ('*le* populisme de droite'), as if its meaning were self-evident. The association of 'populism' with 'right-wing' and its juxtaposition with 'extreme-left' establish a discursive symmetry: populism is positioned along an axis of radicality, operating as a mitigated synonym for extremism and deviation from democratic normality. This parallel implicitly constructs political centrism and moderation as the benchmarks of democratic legitimacy. The convergence of populism with radicalism is further reinforced by the identification of

its cause, namely, the 'radicalisation of the electorate'. The expression 'submerged by right-wing populism' carries a distinctly negative evaluative charge through the metaphor of submersion, which evokes loss of control and collective threat. This metaphor contributes to depoliticising the phenomenon: rather than depicting populism as a social dynamic or a recognisable political discourse, it presents it as an irrational, quasi-natural force capable of engulfing entire countries.

(1) *En attendant, la radicalisation d'une partie de l'électorat pose de sérieuses questions. La Flandre semble suivre la voie d'une série de plus en plus longue de pays européens submergés par le populisme de droite. La Wallonie et Bruxelles font figure d'exception en penchant à gauche mais à la faveur, entre autres, du score du PTB, parti d'extrême-gauche, d'origine maoïste, qui ne s'est jusqu'à présent jamais frotté au pouvoir et n'entend pas dévier.*

In the meantime, the radicalization of part of the electorate raises serious questions. Flanders seems to be following in the footsteps of a growing number of European countries submerged by right-wing populism. Wallonia and Brussels are exceptions, leaning to the left, but this is due in part to the strong results of the PTB, an extreme-left party with Maoist origins, which has never been in power and has no intention of deviating from its course.

(*Metro*, 27 May 2019)

The coding process also required contextual knowledge of the political and media landscapes of Belgium (both Dutch- and French-speaking), France, and Spain. In example (1), understanding the Belgian political context made it possible to infer that the phrase 'far-right populism' refers to the Flemish party Vlaams Belang, even though it is not explicitly mentioned. As the literature underscores, coding frameworks that aim to encode such kind of observations tend to depress ICR scores, precisely because they require interpretive labor at the intersection of text and context (O'Connor & Joffe, 2020: 8). This tension suggests that reduced statistical agreement should not be read simply as coder inconsistency, but rather as an index of the epistemological difficulty of capturing discursive phenomena that operate precisely through ambiguity and semantic drift. To support coder judgment without overly constraining it, our codebook includes concrete examples and explanatory notes, which proved useful in clarifying borderline cases, though the inherently interpretive nature of the task left room for reasonable disagreement.

An additional methodological issue arises from the fact that multiple codes were allowed. Coders could assign several codes to the same unit when appropriate (see below). For example, one coder might classify

an occurrence of *populis** as part of both elite and nativist critique (see Hambye *et al*., Chapter 3 in this volume), while another might code it only as elite critique. This raises important questions about how to evaluate coder agreement when partial overlap, rather than complete alignment, is expected. Standard ICR metrics generally assume binary or mutually exclusive coding and therefore struggle to accommodate justified plurality. For coding schemes that reflect the inherent ambiguity and multidimensionality of public discourse, such measures risk underestimating substantive agreement. As we discuss below, we argue that, in these contexts, agreement should be reconceptualised in terms of substantive convergence rather than strict formal equivalence, that is, whether coders identify the same core discursive function, even if one also recognises additional interpretive layers.

3 Description of the intercoder reliability procedure

As the study sought to capture not only formal textual properties but also implicit meanings, we designed a reliability strategy attentive to both statistical rigor and the epistemic challenges of interpretive coding. Our approach was shaped by three considerations. First, coders needed systematic training and calibration to ensure a shared understanding of the categories, which range from strictly formal descriptors to conceptually demanding interpretative judgments. Second, the interdisciplinary composition of the coding team – spanning linguistics, political science, and communication science – required procedures that accommodate diverse analytical perspectives while still enabling robust measures of convergence. Third, the coding grid itself integrates multiple category formats (closed, open, and mixed), which required different reliability metrics and harmonisation strategies. In what follows, we describe the composition and training of the coding team, the sampling design for the reliability test, and the procedures used to calculate and interpret agreement across the four category types.

3.1 Coding grid, coders, and coding categories

The codebook, developed by one of the principal investigators and presented in Chapter 3, was designed with a dual purpose: to capture both the formal properties of discourse and the implicit dimension. Accordingly, the coding grid integrates closed, open, and mixed category types (see Table 4.1), with variation in the number of admissible responses. In practice, the coding scheme combines factual descriptors, such as speaker identity and institutional arena, with interpretative dimensions that address implicit meanings, including the discursive function of *populis** within the

Table 4.1 ICR method per type of coding category

Type	*Description*	*ICR method*	*Example of categories*
A	Closed list, single response	Krippendorff's alpha	Addressee type, Discursive genre, Function, Grammatical category, Object type, Reference to type, Reported speech, Reported speech source type, Self-evident meaning, Speaker type, Value, Target type, Type of marking
B	Closed list, multiple responses allowed	Krippendorff's alpha	Conceptual field
C	Open set of responses	% agreement	Associated terms, Object ID, RS Author ID, RS Source ID, Speaker ID
D	Mixed open/ closed, multiple responses	% agreement	Addressee ID, Reference to ID, Target ID

message, the conceptual field in which it is embedded, and its value. This, in turn, required broad theoretical grounding, drawing on key concepts, such as extremism, elite critique, and nationalism.

To ensure the methodological robustness of the coding process, particular emphasis was placed on the recruitment, training, and evaluation of coders involved in establishing ICR. Coders were selected from students in the three disciplinary domains represented in the project, that is, linguistics, political science, and communication science. Unlike many qualitative research projects (Campbell *et al.*, 2013), the project thus incorporated multiple coders from the very beginning. This team-based design made it possible, at a later stage, to conduct an ICR assessment without additional recruitment (i.e., external to the coding team). Given the conceptual sophistication of the coding frame, a carefully staged training protocol was implemented. The process began with a three-hour introductory seminar led by one of the principal investigators and the primary architect of the coding grid. During this session, the full codebook was presented and exemplified through detailed examples, and special instructions were introduced to address recurrent edge cases.

Following the introductory session, each coder independently completed a trial coding exercise consisting of 15 occurrences, deliberately selected for their complexity and because they had previously prompted disagreement or discussion within the research team during an earlier test phase. The performance of the coders was then assessed by the project's core team (principal

investigators and researchers), who provided written and verbal feedback. A second three-hour training session was held to review these initial results. This session aimed at resolving coding ambiguities and reinforcing interpretive consistency through group discussion and targeted clarification. After this stage, each coder conducted an initial round of coding on a subset of 80 occurrences. This work was subsequently double-checked by a member of the research team, not with the intention of systematically correcting every potential error as if a single correct interpretation existed, but to identify and address recurring divergences between the coder's application of the codebook and the practices expected by the team. This comparative review served to detect possible misunderstandings of the coding framework and to further calibrate individual interpretations to ensure coherent use of the analytical categories across the research team.

The category system, summarised in Table 4.1, comprises four recurring types that also structured the strategy for calculating ICR (see below). Type A categories are closed lists with a single permitted response. They include several variables – amongst which notably the grammatical form of *populis** (see Appendix 4.1 for a thorough description of the variables), the discursive genre of the source text, the type of speaker employing *populis** (e.g., journalist, politician, expert, other), whether it appears in direct or reported speech, as well as the valence and function of the occurrence. Type B consists of a single category, the conceptual field associated with *populis** (e.g., nationalism, extremism, Euroscepticism, demagoguery), coded as a closed list with multiple possible responses. Type C categories are open-ended and allow for a wide range of responses, including the specific terms associated with populism in the discourse, the identity of the speaker using *populis**, or the object to which *populis** refers. Finally, Type D combines closed and open formats and permits multiple responses; examples include identifying the target of the message containing *populis**, that is, the person, group, or institution explicitly targeted or criticised through the use of *populis**, or the intended addressee.

3.2 ICR strategy

ICR was assessed through a 'double-coding' procedure, in which coders independently re-coded data segments previously coded by one of their peers. To preserve the independence of judgment, this was done without any prior discussion or coordination between coders. Only a portion of the dataset was double coded for ICR because of resource constraints. The subset comprised 300 items, representing 10% of the full corpus. These 300 items were randomly selected from the corpus already coded by the six project coders (designated as Coders 1 in Table 4.2), with 50 items drawn from each coder's contributions. Each batch of 50 items was subsequently

divided into two subsets of 25, which were then independently re-coded by two colleagues from within the same pool of six (designated as Coders 2 in Table 4.2). This procedure yielded a total of ten unique coder pairings, rather than the six that would have resulted from simple one-to-one pairing, thereby enhancing the diversity of combinations. The full allocation matrix is presented in Table 4.2. In total, the process produced 300 double-coded segments (286 after a cleansing phase) spanning ten sub-corpora, increasing not only the robustness of the reliability assessment but also the variation in coder interactions. This expanded diversity may reduce raw agreement scores but offers a more realistic measure of reliability across epistemic positions. At the same time, the broader distribution of pairings helps dilute pair-specific effects and avoids strongly skewed agreement patterns driven by chance or by atypical coder combinations, reducing the influence of outliers and producing a more balanced estimate of overall reliability.

As shown in Table 4.1, reliability metrics were matched to category type. For closed list, single response categories, Type A, intercoder agreement was evaluated using Krippendorff's alpha. For closed list, multiple response categories, Type B, Krippendorff's alpha was also used with adaptations appropriate for multi-label coding. Our choice of Krippendorff's alpha was motivated by its suitability for datasets with multiple coders, missing values, and mixed formats. Krippendorff's alpha computes agreement across the full set of coders and the entire category, which makes it efficient and scalable in a design like ours and appropriate for (Feng, 2013: 2960). Cohen's kappa is frequently used in reliability studies, yet it requires pairwise comparisons and is therefore less efficient in multi-coder (Hruschka *et al.*, 2004: 313). Krippendorff's alpha corrects this inflation and provides a more conservative and realistic estimate. We interpreted

Table 4.2 ICR distribution

Coder 1	*Coder 2*	*N items*
A	B	25
	C	25
B	C	25
	D	25
C	E	25
	F	25
D	A	25
	E	25
E	A	25
	F	25
F	B	25
	D	25

(e.g., Cicchetti, 1994; Landis & Koch, 1977) where values above 0.80 are often described as almost perfect or excellent, and values below 0.40 are often described as poor. At the same time, we followed O'Connor and Joffe (2020: 9), in treating such thresholds as indicative rather than binding. Benchmarks remain arbitrary unless situated in the context of the task, the distribution of categories, and the consequences of coding errors, and we therefore contextualised any reported alpha accordingly.

For open answer categories, Type C, and for mixed categories that combine open and closed response options with multiple possible answers, Type D, we relied on manual harmonisation procedures and percentage agreement used cautiously, with explicit acknowledgement of the limits of simple agreement measures to account for chance agreement. While the choice to include open and mixed categories was conceptually motivated by the multidimensional character of *populis** and of its discursive functions (Hambye *et al.,* Chapter 3 in this volume), it also had computational consequences. In these categories, chance agreement in the strict statistical sense was practically impossible, which removed the rationale for some standard reliability estimators (Enschot *et al*., 2024: 5) and therefore justified relying on simple agreement measures.

A central methodological challenge concerned the evaluation of coder convergence in categories B, C, and D when responses were either not identical or downright entirely divergent. To address this, we introduced a typology of intermediate correspondence that captures partial alignments, as shown in Table 4.3. First, partial agreement with extent divergence occurs when one coder lists more relevant items than the other but at least one item matches, for example, Coder 1 codes Viktor Orbán and Donald Trump and Coder 2 codes Orbán only. Second, partial agreement with composition divergence arises when the number of items is similar, but the referents are conceptually related rather than identical, for example, Salvini versus Lega Nord (Salvini's political party). Third, partial agreement

Table 4.3 Types of agreement for categories BCD

Type of agreement	*Coder 1*	*Coder 2*
Perfect agreement	Orbán	Viktor Orbán
Partial agreement – extent divergence	Orbán; Trump	Viktor Orbán
Partial agreement – composition divergence	Donald Trump; Brexit, M5S; Lega Nord; Viktor Orbán	Trump; Salvini; Orbán; Johnson
Partial agreement – extent-composition divergence	Orbán; Salvini; Lepen	Orbán; Trump; Johnson
Disagreement	Orbán	Trump

with extent-composition divergence appears when some items coincide while others diverge in both form and substance, for example, Orbán, Salvini, Le Pen versus Orbán, Trump, and Johnson. Distinguishing the type of agreement as such adds value to the coding process by making interpretive discrepancies analytically visible instead of treating them as mere noise. It allows coders to document and justify divergent interpretations while maintaining a shared analytical framework. From the perspective of ICR, it introduces a graduated understanding of convergence, going beyond the binary logic of agreement/disagreement that often obscures the interpretive dimension of qualitative coding. In the analysis of results, this approach provides a fine-grained view of how meaning is negotiated through the coding process. It helps identify systematic patterns of divergence – whether stemming from conceptual overlap, contextual ambiguity, or differential salience judgments – and thus informs subsequent discussions about category refinement and coder calibration.

An additional consideration was that certain dimensions in the coding grid follow a hierarchical logic. Coders were first asked to identify an overarching category, for example, the speaker using *populis**, and only then to specify a subtype, for example, journalist, politician, and expert. Following van Enschot *et al.* (2024: 13), we calculated agreement for subcategories only when coders concurred on the parent category. This contingent approach avoids distortions in reliability metrics that would otherwise arise from upstream disagreement (Enschot *et al.*, 2024: 11).

4 Interpreting the ICR results

As in many interpretative research projects (Hruschka *et al.*, 2004: 321), our initial round of double-coding produced relatively low agreement scores (see Appendix 4.2). Rather than interpreting this outcome as a methodological shortcoming, we approached it as an analytic opportunity: a chance to probe more deeply into the sources of divergence between coders and to reflect on what such divergences reveal about the practice of assessing ICR in conceptually rich, interpretative inquiry.

The purpose of this section is twofold. First, we present the main empirical patterns of coder agreement and disagreement, distinguishing between variables that yield high reliability and those that prove more challenging. This descriptive synthesis highlights how category design, distributional properties, and the nature of the interpretative task shape the outcomes of statistical measures such as Krippendorff's alpha. Second, we move beyond description to consider what these patterns reveal about the practice of assessing reliability in interpretative research. We argue that statistical

indices provide only a partial picture: they must be interpreted in light of procedural factors (such as coder training and coding rules), individual factors (such as coders' familiarity with the sociopolitical discourse under study and their willingness to engage rigorously and conscientiously with the interpretative task), and epistemological choices (such as whether categories are open, closed, or deliberately polysemic).

The ICR assessment covers 19 variables spanning the four main category types of the coding grid. Overall, the results reveal a clear divide between formal descriptors and interpretative dimensions, with the former yielding consistently higher levels of agreement. Type A variables, which required coders to select a single response from a closed list, show the clearest internal differentiation. Grammatical and structural descriptors perform best: GRAMMATICAL CATEGORY and OBJECT TYPE display strong Krippendorff's alpha coefficients (0.83 and 0.91, respectively), indicating consistent identification of the morphosyntactic form of *populis** and the ontological type of its referent. SPEAKER TYPE also performs well ($\alpha = 0.78$), showing broad coder consensus on whether the utterance originated from a politician, journalist, or citizen. By contrast, interpretative dimensions such as DISCURSIVE GENRE ($\alpha = 0.41$), VALUE ($\alpha = 0.35$), and FUNCTION ($\alpha = 0.41$) yield lower reliability. These categories demand contextual judgment – determining, for instance, whether *populis** is used descriptively or evaluatively, thus inviting interpretive variation. REPORTED SPEECH and REPORTED SPEECH SOURCE TYPE both achieve moderate reliability ($\alpha = 0.66$), as coders occasionally differed in identifying whether the term appeared in direct quotation or verbatim. SELF-EVIDENT MEANING represents a special case: raw agreement is extremely high (97%), yet alpha is modest (.54) due to a skewed category distribution, with nearly all instances coded as 'Yes'. Similarly, REFERENCE TO TYPE, ADDRESSEE TYPE, and TYPE OF MARKING (α between 0.32 and 0.41) perform weakly, confirming that inferential tasks, such as identifying communicative relations or evaluative mechanisms, reduce convergence. TARGET TYPE ($\alpha = 0.47$) follows this pattern, as coders sometimes disagreed on whether the implicit target was an individual actor, an institution, or a broader collective like 'the elite'.

This divide between formal and interpretive dimensions becomes even more pronounced when moving from single-response closed categories (Type A) to multi-response or open formats (Types B, C, D). In these latter categories, coders had to not only recognise explicit textual features but also reconstruct semantic and pragmatic relations around *populis** based on lexical cues. CONCEPTUAL FIELD, the only Type B variable, exemplifies this dynamic. With an agreement rate of 43%, it yields one of the lowest reliability scores. Coders typically agreed on one salient conceptual domain, such as demagogic, but diverged on whether additional dimensions (anti-elite, nationalist, conspiratorial) were also relevant. This pattern illustrates how multi-label coding captures interpretive richness at the

cost of statistical convergence. Yet, as shown by the partial agreement rate (see Appendix 4.2), such cases often reflect substantive convergence on a shared interpretive core, even when coders differed in scope or emphasis.

Type C variables, which allow open-ended responses, produce more heterogeneous but generally acceptable outcomes. SPEAKER ID and OBJECT ID achieve moderate to good agreement (73% and 61%), indicating frequent convergence on referents. ASSOCIATED TERMS generates the highest proportion of partial agreements (32%): coders often shared at least one term linked to *populis**, though their lists rarely matched exactly. REPORTED SPEECH AUTHOR ID and REPORTED SPEECH SOURCE ID reach higher reliability (87% and 80%), as these categories rely on factual identification once coders agreed on the presence of reported speech. Type D variables, which combine open and closed formats with multiple possible responses, yield both the strongest and weakest results. ADDRESSEE ID and REFERENCE ID show very high agreement (94% and 92%), confirming coders' consistency in identifying the interlocutor and intertextual reference of an utterance. TARGET ID, however, displays low reliability (57%), reflecting the interpretive difficulty of discerning implicit or indirect targets in complex discourse segments.

Taken together, the ICR results reveal a consistent pattern. Variables anchored in clear, directly observable textual features, such as GRAMMATICAL CATEGORY, OBJECT TYPE, ADDRESSEE ID, and REFERENCE ID, achieve strong convergence, as their definitions are operationalised through mutually exclusive categories supported by examples in the codebook. As a result, convergence is consistently high, with Krippendorff's alpha values above the conventional 0.80 threshold in most cases. Conversely, categories requiring inferential or contextual interpretation, such as VALUE, FUNCTION, OR CONCEPTUAL FIELD, produce weaker agreement. Their operational difficulty stems from conceptual abstraction, overlapping meanings, and the need to weigh tone, irony, or evaluative framing. A low Krippendorff's alpha, however, does not imply that coders disagreed on the overall tendency – for example, that the VALUE of most occurrences of *populis** is negative. The coefficient reflects the absence of strong convergence in coding decisions, not unreliability in the underlying data or in the interpretive conclusions drawn from it.

Additional complexity arises from structural and statistical factors. Multi-label categories inherently broaden the interpretive space, while variables with highly skewed distributions, like VALUE or SELF-EVIDENT MEANING, inflate expected chance agreement and depress alpha coefficients. From this perspective, low alpha does not necessarily indicate error, but the open-ended nature of interpretive work itself. Krippendorff's alpha, which adjusts for agreement expected by chance, is designed precisely to reveal this: it measures the degree of interpretive convergence among coders across all possible categories, not the likelihood of coding error in the results (which remains low). DISCURSIVE GENRE poses a distinct but related difficulty: the classification depends partly on the

coder's background knowledge of communicative forms and media conventions, which varies across individuals. Such variation in familiarity produces uneven applications of genre distinctions, leading to lower reliability scores.

Finally, partial agreements, especially in CONCEPTUAL FIELD, carry analytical significance. They reflect substantive convergence on core meaning accompanied by variation in interpretive breadth. Such cases mark zones of semantic density and contestation. The divergences that can be found therefore reflect differences in interpretive background and analytical emphasis: distinct disciplinary lenses or theoretical sensitivities can lead coders to privilege different semantic associations, all justifiable within the coding frame. TARGET ID displays a similar pattern. Because targets are not always explicitly named, coders frequently had to infer them from their knowledge (see (1)), reconstructing the implied object of critique or reference. This task, by nature, invites variation in interpretation.

Overall, coder divergence emerges in predictable and interpretable ways. Variables anchored in directly observable textual features tend to produce high scores, whereas those requiring coders to weigh contextual cues, infer implicit meanings, or prioritise among competing interpretive possibilities yield lower scores. This pattern, namely lower convergence on interpretive categories, has also been reported in previous work on populism (see Hatakka & Herkman, 2022: 1528). Structural characteristics such as multi-label coding or skewed distributions further complicate the statistical picture, sometimes depressing coefficients despite significant substantive alignment. These findings underscore the need to interpret ICR statistics with care – especially in research explicitly designed to capture the polysemic, layered, and interpretive character of political discourse.

At the same time, the ICR exercise proves methodologically productive beyond its diagnostic value. It prompts a reflexive refinement of the coding scheme and encourages cautious treatment of ambiguous cases. In several instances, disagreement reveals that certain categories are intrinsically difficult to apply because the discourse itself is ambivalent. Flagging such borderline cases proved useful for two reasons: first, it allowed us to decide whether to include, exclude, or explicitly annotate them in quantitative summaries; and second, it identified analytically rich examples for subsequent, more qualitative explorations. In this sense, ICR does not merely assess the convergence of coding decisions. It also serves as a device for detecting zones of ambiguity and interpretive interest within the corpus.

5 Considerations for assessing reliability in interpretive research

The descriptive patterns outlined above show that ICR in interpretive research, especially discourse analysis, is not reducible to technical indices. To understand why some categories achieve high convergence while

others generate disagreement, it is necessary to situate reliability outcomes within a wider framework. Based on what we observed, three domains, procedural and technical, contextual, and epistemological, jointly shape reliability in ways that conventional metrics only partially capture.

5.1 Procedural and technical factors

The first layer of influence stems from the mechanics of coding itself. Coders in this project received training through seminars, trial exercises, and feedback loops, which helped standardise their application of the scheme. Nonetheless, some limitations persisted. For example, coders who were not involved in the original design of the codebook found it harder to fully internalise its logic, which occasionally resulted in lapses such as leaving lines uncoded, inconsistently applying the 'x' marker for inapplicable categories, or discontinuing their assignments. These issues echo the literature on naïve or insufficiently trained coders, which highlights the persistent gap between formal comprehension of coding instructions and their consistent operational execution in complex interpretative tasks, in spite of the training module they attended (Enschot *et al*., 2024: 2; Spooren & Degand, 2010).

Technical properties of the data also play a significant role. As noted above, variables with skewed distributions, such as VALUE or SELF-EVIDENT MEANING, produce low coefficients. Krippendorff's alpha, like Cohen's kappa, is sensitive to prevalence effects: when one category dominates, expected chance agreement increases, thereby lowering the coefficient even if coders substantively agree in most cases (Di Eugenio & Glass, 2004; Jeni *et al*., 2013). This effect can however be mitigated through the use of alternatives such as Kappa Max (Umesh *et al*., 1989), which adjusts observed kappa relative to the maximum possible given the data's marginal distributions, and PABAK (Byrt *et al*., 1993). While rarely applied in qualitative research, these adjustments underscore the need to pair numerical indices with awareness of their limits.

5.2 Human factors

A second domain concerns the coders' familiarity with the sociopolitical and cultural context of the material. Because the corpus included discourse from Belgium, France, and Spain, coders with limited prior exposure to these contexts faced difficulties in interpreting certain idioms, references, or implicit cues. This sometimes led to misidentification of referents or under-detection of implied meanings. For instance, judgments on TARGET ID often hinge on contextual knowledge, such as inferring which political actor is implied as the target. In such cases, divergence reflected uneven contextual competence. Beyond contextual knowledge, this points to a broader human or individual factor – the degree of professionalism and commitment that coders bring to the interpretative task. A professional

and committed coder, aiming for a 'job well done', will actively seek to understand what their contextual knowledge alone does not allow them to grasp, instead of settling for surface interpretation. This engagement dimension also has concrete methodological implications: for example, the initial distribution of coder pairs (see Table 4.2) had to be revised following the withdrawal of one participant before completing the assigned task.

In this sense, differences in reliability may also depend on varying degrees of diligence and involvement. It is then the combination of contextual knowledge and professionalism that allows bridging the gap between abstract category definitions and the situated meanings embedded in discourse.

5.3 Epistemological factors

The third and most profound layer of influence stems from the epistemological orientation of the coding scheme. Categories in Types B, C, and D were deliberately designed to capture polysemy, overlap, and multiplicity – features intrinsic to political discourse. This openness was an intentional design choice, in line with the aim of the project to preserve the layered character of discourse rather than reduce it to a single 'correct' code. The correlation between conceptual complexity and lower reliability is well documented in the literature (Feng, 2013: 1804; Fort *et al.*, 2012: 898), and our results confirm this: interpretative categories with broad or polysemic definitions return lower alpha values not because of coder error, but because multiple defensible judgments are possible.

6 Conclusion

The analysis presented in this chapter demonstrates that reliability in qualitative research cannot be meaningfully reduced to numerical indices. While statistical measures such as Krippendorff's α provide a useful snapshot of coder convergence, their interpretive value depends fundamentally on the nature of the variables under examination. In the present study, high agreement on formal descriptors and lower convergence on interpretative dimensions reflect not methodological inconsistency but the very nature of the analytical task: the attempt to capture implicit, polysemic, ambiguous, and context-dependent meanings. Beyond mere scores, the ICR exercise reveals that coder divergence can itself be analytically productive. When properly documented and discussed, disagreement exposes zones of semantic ambiguity, conceptual tension, or contextual complexity that warrant closer examination. Rather than being treated as error or noise, such divergences point to the discursive density of the material and guide the researcher toward passages of particular interpretive interest.

The results underscore the importance of coder training, contextual competence, and professional engagement. Familiarity with the sociopolitical environments (Belgium, France, and Spain in our case) proves essential for consistent interpretation, while differences in diligence or depth of engagement shape the precision and granularity of coding decisions. At a more fundamental level, the findings reaffirm the epistemological specificity of interpretive reliability. Categories designed to capture polysemy, implicit meaning, or evaluative nuance will, by design, produce lower statistical convergence. This outcome reflects a research philosophy that privileges meaning reconstruction over mechanical replication. The challenge is therefore to make the interpretive process traceable, transparent, and theoretically coherent.

References

Braun, Virginia, & Clarke, Victoria. (2013). *Successful qualitative research: A practical guide for beginners*. SAGE.

Byrt, Ted, Bishop, Janet, & Carlin, John B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-v

Campbell, John L., Quincy, Charles, Osserman, Jordan, & Pedersen, Ove K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Carmines, Edward, & Zeller, Richard. (1979). *Reliability and validity assessment*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985642

Cicchetti, Domenic V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cofie, Nicholas, Braund, Heather, & Dalgarno, Nancy. (2022). Eight ways to get a grip on intercoder reliability using qualitative-based measures. *Canadian Medical Education Journal / Revue Canadienne de l'éducation Médicale*, *13*(2), 73–76. https://doi.org/10.36834/cmej.72504

Di Eugenio, Barbara, & Glass, Michael. (2004). Squibs and discussions: The Kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. https://doi.org/10.1162/089120104773633402

Feng, Guangchao C. (2013). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality & Quantity*, *47*(5), 2959–2982. https://doi.org/10.1007/s11135-012-9745-9

Fort, Karen, Nazarenko, Adeline, & Rosset, Sophie. (2012). Modeling the complexity of manual annotation tasks: A grid of analysis. In Martin Kay & Christian Boitet (Eds.), *Proceedings of COLING 2012* (pp. 895–910). The COLING 2012 Organizing Committee. https://aclanthology.org/C12-1055/

Hambye, Philippe, & Rondiat, Coline. (2024). *Sociopolitical keywords in contemporary discourse: Addressing methodological challenges for qualitative approaches*. MethodsNET Launch Events, Louvain-la-Neuve.

Hatakka, Niko, & Herkman, Juha. (2022). Hegemonic meanings of populism: Populism as a signifier in legacy dailies of six countries 2000–2018. *Media, Culture & Society*, 44(*8*), 1523–1540.

Hruschka, Daniel J., Schwartz, Deborah, St. John, Daphne C., Picone-Decaro, Erin, Jenkins, Richard A., & Carey, James W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods*, *16*(3), 307–331. https://doi.org/10.1177/1525822X04266540

Jeni, László A., Cohn, Jeffrey F., & De La Torre, Fernando. (2013). Facing imbalanced data recommendations for the use of performance metrics. *International Conference on Affective Computing and Intelligent Interaction and Workshops: [Proceedings]. ACII (Conference), 2013*, 245–251. https://doi.org/10.1109/ACII.2013.47

Krippendorff, Klaus. (1980). *Content analysis: An introduction to its methodology* (2nd ed., [Nachdr.]). Sage Publications.

Landis, Richard J., & Koch, Gary G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Miles, Matthew B., & Huberman, A. Michael. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications.

O'Connor, Cliodhna, & Joffe, Helene. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1–13.

Spooren, Wilbert, & Degand, Liesbeth. (2010). *Coding coherence relations: Reliability and validity*, *6*(2), 241–266. https://doi.org/10.1515/cllt.2010.009

Stenbacka, Caroline. (2001). Qualitative research requires quality concepts of its own. *Management Decision*, *39*(7), 551–556. https://doi.org/10.1108/EUM0000000005801

Umesh, Uchila N., Peterson, Robert A., & Sauber, Matthew H. (1989). Interjudge agreement and the maximum value of Kappa. *Educational and Psychological Measurement*, *49*(4), 835–850. https://doi.org/10.1177/001316448904900407

van Enschot, Renske, Spooren, Wilbert, van de Bosch, Antal, Burgers, Christian, Degand, Liesbeth, Evers-Vermeul, Jacqueline, Kunneman, Florian, Liebrecht, Christine, Linders, Yvette, & Maes, Alfons. (2024). Taming our wild data: On intercoder reliability in discourse research. *Dutch Journal of Applied Linguistics*, *13*. https://doi.org/10.51751/dujal16248

Wilson-Lopez, Amy, Minichiello, Angela, & Green, Theresa. (2019). An inquiry into the use of intercoder reliability measures in qualitative research. *2019 ASEE Annual Conference & Exposition Proceedings*, Tampa, Florida. https://doi.org/10.18260/1-2--32067

Appendix 4.1 – Codebook variables

Example: *Et après un appel à la population à «faire bloc» #Macron ose encore accuser les autres de faire dans le #populisme et la démagogie 🗣🤡💅Les Français ont élu un #gouvernement pour gérer les affaires du pays et non pas pour que le contribuable😇fasse de la police de proximité!😒*

And after calling on the population to "stand united," #Macron dares to accuse others of engaging in #populism and demagogy 🗣🤡💅 The French elected a #government to run the country's affairs, not so that taxpayers 😇 would do community policing! 😒

Variable	*Description*	*Example*
Grammatical category	Indicates the grammatical form of the occurrence of populis* in context. Possible values: adjective (epithet or attributive), definite noun, or indefinite noun. This identifies the morphosyntactic status of the term.	Definite noun
Associated terms	Lists the term(s) syntactically or semantically linked to populis*, such as modifiers, complements, or coordinated items. If multiple, separate by semicolon (';'). If none, code X.	Demagogy
Discursive genre	Refers to the type of discourse or text in which the occurrence appears, according to institutional and communicative conventions. For media data, options include news brief, report, editorial, analysis, interview, and opinion piece. For parliamentary data, options include oral intervention, written communication, question to government, and report.	Tweet
Speaker ID	Unique identifier or name of the speaker who produced the utterance (the person responsible for the use of populis*). Can be a proper name, pseudonym, or Twitter handle.	Lambda twitter user

(*Continued*)

Variable	*Description*	*Example*
Speaker type	Identifies the social or professional status of the speaker (the person who produced the utterance containing populis*). Possible values: politician, journalist/press agency, expert, columnist, public figure, other, Unknown	X
Reported speech	Indicates whether the term populis* appears as part of another person's discourse (i.e., whether the speaker is quoting, paraphrasing, or reproducing someone else's words). Possible values: No / Verbatim / DR (reported discourse).	Reported discourse
RS Author ID	Identifies the author of the reported speech (the person whose words are quoted or paraphrased when populis* occurs).	Emmanuel Macron
RS Source ID	Identifies the source or medium (e.g., Le Monde, European Parliament, Twitter) from which the reported speech originates.	X
RS Source type	Identifies the original arena or type of medium from which the reported speech originates: political, traditional media, social media, other.	X
Object ID	Identifies the concrete entity (individual, group, action, etc.) to which the attribute populist is applied in the utterance. If self-referential, code AR.	Self-referential
Object type	Specifies the ontological category of the entity characterised as populist (the object of the characterisation). Values include individual, defined group, undefined group, individual/group, action/event/fact. If the use is self-referential, leave blank or code 'X'.	X
Value	Captures the evaluative polarity attached to populis* in the utterance. Possible values: positive, positive-ironic, negative, negative-ironic, neutral, ambivalent, NC. This assesses the explicit or implicit normative charge of the term.	Negative
Type of marking	Indicates the linguistic mechanism by which the evaluative value of populis* is marked. Values: lexical (value-bearing adjective, noun, or verb), person/group (contrast or association with a valued referent), inference. If Value is neutral or NC, leave blank.	Lexical

(*Continued*)

Variable	*Description*	*Example*
Self-evident meaning	Indicates whether the meaning of populis* is presented as self-evident or as requiring clarification. Yes means the meaning is taken for granted (no definition provided). No means the meaning is questioned, explained, or defined in context.	Yes
Conceptual field	Identifies the broader conceptual domain or ideological frame associated with populis* in context. Possible values: demagogic, anti-system/anti-elite, traditionalist/ conservative, nationalist/ethnicist/ souverainist, popular/coarse, extremist/ radical, autocratic, conspiratorial/dishonest, people-based, NC.	Demagogy
Function	Refers to the primary pragmatic or argumentative role of the occurrence within the utterance. Possible values include description/characterisation, critique of process, critique of act, critique of actor/ group, denial, legitimisation by opposition, (self-)legitimisation/promotion.	Description
Target ID	Identifies the person, group, or institution explicitly targeted or criticised through the utterance (when applicable).	Macron
Target type	Refers to the social or institutional type of the target explicitly criticised or attacked through the use of populis* (distinct from the grammatical object). Uses the same typology as Speaker Type (politician, journalist, expert, etc.).	Politician
Reference ID	Identifies the external message or source text to which the utterance refers (if the occurrence comments on another text or message).	X
Reference to Type	Identifies whether the utterance is primarily a commentary or reaction to another message or text.	X
Addressee ID	Identifies the person or group to whom the utterance is directly addressed (e.g., through mention, response, or reply).	X
Addressee type	Specifies whether the utterance is directly addressed to another actor, and if so, what form the address takes.	X

Appendix 4.2 – ICR results

Variable	*Type*	*Total items*	*N Disagreement*	*N Agreement*	*N Partial agreement*	*Agreement rate*	*Partial agreement rate*	*Krippendorff's α*	*Agreement type*
Grammatical category	A	280	28	252		0.9		0.831	Strong
Discursive genre	A		111	120		0.519		0.406	*Poor*
Speaker type	A	231	29	171		0.855		0.78	Acceptable
Object type	A	200	11	190		0.945		0.906	Strong
Target type	A	201	2	17		0.895		0.475	*Poor*
Reported speech	A	19	55	229		0.806		0.661	*Poor*
RS source type	A	284	2	5		0.714		0.664	*Poor*
Value	A	7	86	195		0.694		0.346	*Poor*
Self-evident meaning	A	281	8	276		0.972		0.542	*Poor*
Function	A	284	120	159		0.57		0.405	*Poor*
Reference to type	A	279	114	164		0.573		0.354	*Poor*
Addressee TYPE	A	278	19	267		0.934		0.282	*Poor*
Type of marketing	A	286	111	175	0	0.612	/	0.324	*Poor*

Conceptual field	B	286	133	122	31	0.427	0.108	0.424	*Poor*
Associated terms	C	286	20	169	92	0.590	0.321		*Poor*
Speaker ID	C	286	70	208	8	0.727	0.028		Acceptable
Object ID	C	286	77	173	36	0.605	0.126		Acceptable
RS Author ID	C	286	36	248	0	0.87	/		Acceptable
RS Source ID	C	286	51	229	6	0.8	0.021		Acceptable
Target ID	D	286	114	164	8	0.573	0.028		*Poor*
Addressee ID	D	286	16	268	2	0.937	0.007		Strong
Reference ID	D	286	21	264	6	0.923	0.021		Strong

Agreement type is based on Krippendorff's alpha when applicable, and on the raw agreement rate in other cases. The thresholds follow Krippendorff (2004): ≥ 0.800: Meets standard (reliable); 0.667–0.799: Tentatively acceptable; < 0.667: Below standard (unreliable).



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com