Pavle Savković, BSc

Beyond the Speech: Context-Aware Topic Labelling and Linguistic Style in Parliamentary Debates

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme: Computational Social Systems

submitted to

Graz University of Technology

Supervisor

Dipl.-Wirt.-Inf. Dr.rer.soc.oec. Stefan Thalmann
Institute for Operations and Information Systems

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helić
Institute of Human-Centred Computing (HCC)

Graz, November 2025

Abstract

Parliamentary debates are turn-taking, dialogue-like conversations, yet most large-scale text analysis treats speeches as isolated documents. We introduce a sequence/context aware pipeline for topic classification and stylistic profiling that respects the dialog structure of parliamentary discourse. Using 1.4M speeches from Austria, Great Britain, and Croatia (1996–2022) in the ParlaMint corpus, we first segment debates into coherent agenda episodes, then embed segments, cluster them, and map unsupervised topics to Comparative Agendas Project (CAP) domains. In parallel, we profile linguistic style with LIWC-22.

Our segment-first design prioritises episode coherence over per-utterance precision, providing interpretable labels suitable for temporal analysis and policy monitoring. Benchmarks against ParlaCAP automatic labels and human test sets indicate sufficient reliability for context-preserving analysis at the episode level. The method reveals stable cross-national patterns: in Macroeconomic debates we observe more transactional and money-related vocabulary and less moral/insight language; coalition speakers use more positive tone and collective pronouns than opposition; and crises produce predictable agenda substitutions (e.g., Health rises during COVID-19 as Macroeconomics falls).

By treating debates as sequences rather than bags of speeches, we show that institutional position and policy domain systematically shape linguistic style more than individual or party characteristics, while enabling robust, interpretable tracking of issue attention over time.

Kurzfassung

Parlamentsdebatten sind abwechselnde, dialogartige Gespräche, doch die meiste großskalige Textanalyse behandelt Redebeiträge als isolierte Dokumente. Wir stellen eine sequenzbewusste Pipeline für Themenklassifikation und stilistische Analyse vor, die die dialogische Struktur parlamentarischer Kommunikation respektiert. Auf Basis von 1,4 Mio. Reden aus Österreich, Großbritannien und Kroatien (1996–2022) im ParlaMint-Korpus segmentieren wir Debatten zunächst in kohärente Agenda-Episoden, ordnen anschließend unüberwachte Cluster den Domänen des Comparative Agendas Project (CAP) zu und profilieren den sprachlichen Stil mit LIWC-22.

Unser Segment-first-Ansatz priorisiert Segmentkohärenz gegenüber Beitragsgenauigkeit und liefert interpretierbare Labels für Zeitreihenanalysen und Politikmonitoring. Benchmarks deuten auf ausreichende Zuverlässigkeit für konservierende Kontextanalyse auf Episodenebene hin. Die Methode offenbart stabile länderübergreifende Muster: In makroökonomischen Debatten beobachten wir mehr transaktionales Vokabular und weniger Moral-/Einsichtssprache; Redner:innen der Regierungsparteien verwenden einen positiveren Ton und mehr kollektive Pronomen als die Opposition; und Krisen führen zu vorhersehbaren Themenschiebungen (z. B. steigt *Health* während COVID-19, während *Macroeconomics* zurückgeht).

Indem wir Debatten als Sequenzen statt als „Beutel einzelner Reden“ behandeln, zeigen wir, dass institutionelle Position und Politikdomäne den sprachlichen Stil stärker prägen als individuelle oder parteispezifische Merkmale und dass sich damit eine robuste, interpretierbare Verfolgung der Themenaufmerksamkeit über die Zeit ermöglichen lässt.

Acknowledgements

I would like to thank my supervisors, Denis Helić and Stefan Thalmann, for their guidance and for the many discussions that helped shape this thesis. I also thank the ParlaMint and ParlaCAP teams for maintaining the data infrastructure and for providing access to the human-labelled test sets.

Note on AI-assisted tools

During the writing of this thesis, I used large language models (LLMs) such as ChatGPT, Gemini, and Claude as an assistive tool. They were used to brainstorm alternative phrasings, improve the clarity of English sentences, and debug minor code issues. All research questions, methodological design choices, data processing steps, and interpretations of results are my own. I reviewed, edited, and integrated all AI-assisted text myself, and I remain fully responsible for the content of this thesis.

Contents

Abstract 3

Kurzfassung 5

Acknowledgements 7

1. Introduction 23
- 1.1. Data . . . 25
  - 1.1.1. Corpora . . . 25
  - 1.1.2. Corpora Statistics . . . 25
  - 1.1.3. Multilingual Processing . . . 26
  - 1.1.4. Reference Labels and Human Annotations . . . 26
  - 1.1.5. Dataset Illustration . . . 26
- 1.2. Research Questions and Contributions . . . 28

2. Background and Related Work 31
- 2.1. Parliamentary Corpora and Policy Coding . . . 31
- 2.2. Political Language and Style . . . 32
- 2.3. Political Discourse as a Signal for Markets and Risk Management . . . 33
- 2.4. Topic Modelling with Embeddings . . . 34
- 2.5. Sequential Structure in Discourse . . . 35

3. Methods 37
- 3.1. Pipeline Overview . . . 37
- 3.2. Speech and Segment Embeddings . . . 37
- 3.3. Sequential Segmentation . . . 38
- 3.4. Clustering and CAP Alignment . . . 40
- 3.5. Psycholinguistic Profiling . . . 41
- 3.6. Code and Reproducibility . . . 43

4. Results **45**
4.1. Evaluation of Topic Modelling . . . 45
4.1.1. Task mismatch and evaluation philosophy . . . 45
4.1.2. Scores against reference labels . . . 45
4.1.3. Human agreement as a ceiling . . . 46
4.1.4. Error structure and confusion patterns . . . 46
4.1.5. Country-specific challenges . . . 55
4.1.6. What these scores mean for users . . . 55
4.2. Linguistic Profiles Across Policy Domains and Roles . . . 55
4.2.1. Macroeconomics and Health . . . 55
4.2.2. Coalition vs. Opposition . . . 58
4.2.3. Topic–style interactions across all domains . . . 58
4.3. Ideology, Gender, and Age . . . 62
4.3.1. Ideology . . . 62
4.3.2. Gender . . . 62
4.3.3. Age . . . 62
4.4. Temporal Dynamics of Agenda and Style . . . 66
4.4.1. Party Status Tone . . . 66
4.4.2. Political Rhetoric Over Time . . . 67
4.4.3. Cognitive and Temporal Focus Over Time . . . 68
4.4.4. Economic vs. Health Topics . . . 72
4.4.5. Security and International Affairs . . . 73
4.5. Cross-Lingual Embedding Consistency . . . 74

5. Discussion **77**
5.1. Substantive Mechanisms Behind Stable Patterns . . . 77
5.2. Country-Specific Nuance Without Loss of Comparability . . . 77
5.3. Why Sequence Matters for Interpretation . . . 78
5.4. Reading Effect Sizes in LIWC-22 . . . 79
5.5. Robustness and Design Trade-Offs . . . 79
5.6. Limitations and Implications for Evaluation . . . 80
5.7. Practical Implications for Stakeholders . . . 82
5.8. Where This Leaves Supervised Modelling . . . 83

6. Conclusion **85**

Bibliography **87**

A. Human Agreement on CAP Labels **92**

B. LLM Classification Prompt **93**

C. Additional Temporal Analyses **95**

D. Additional Figures from Section 4.4 (Austria) **99**

E. Additional Figures from Section 4.4 (Croatia) **105**

List of Figures

1. **Analytical workflow.** Speeches (text + metadata) are embedded, segmented into agenda episodes, re-embedded at the segment level, clustered, and mapped to CAP domains via keywording. In parallel, LIWC-22 is computed per speech and aggregated by topic, role, party, demographics, and time. The key design choice is segment-first topic labelling, which preserves debate context for downstream style analysis. 37
2. **Political discourse vs. population norms.** Country-wise LIWC-22 z-scores for parliamentary speech relative to general-population benchmarks from the LIWC-22 Test Kitchen Corpus. Positive cells indicate categories that are over-represented in parliamentary language (e.g., political and power vocabulary), while negative cells indicate under-use. This overview panel provides the reference baseline for all later figures: topic-, role-, ideology-, demographic-, and time-specific effects should be interpreted as deviations from the already distinctive style of parliaments shown here. . 44
3. **Great Britain (GB): model vs. ParlaCAP.** Confusion matrix of predicted CAP domains (x-axis) against ParlaCAP labels (y-axis) for the full corpus; the diagonal indicates agreement. Off-diagonal mass concentrates in conceptually adjacent domains (e.g., welfare–health, macro–commerce), consistent with boundary and mixed-episode cases. 47
4. **Great Britain (GB): model vs. human test labels.** Compared to ParlaCAP, human labels typically reduce ambiguity in boundary speeches, but the same broad structure remains: most off-diagonal mass lies between neighbouring policy areas. This supports the interpretation that many "errors" are boundary disagreements (what the episode is *mainly* about), not failures to identify the general policy region. 48

5. **Croatia (HR): model vs. ParlaCAP.** HR shows the same pattern as GB: misclassifications are concentrated between related domains. In practice, these often correspond to agenda items that naturally mix policy frames (e.g., public administration reforms discussed alongside fiscal implications), where segment-level labels prioritise episode coherence. . . . 49
6. **Croatia (HR): model vs. human test labels.** Agreement improves relative to ParlaCAP, but remaining disagreements again cluster around adjacent categories. This suggests that our pipeline preserves the correct policy neighbourhood even when it differs on fine-grained boundaries, which is the expected trade-off when enforcing segment-level consistency. . 50
7. **Austria (AT): model vs. ParlaCAP.** Austria has no human test set in our evaluation, so we report agreement against ParlaCAP on the full corpus. The same qualitative structure holds: disagreements are dominated by adjacent or thematically overlapping CAP domains. This is consistent with the interpretation that differences often reflect alternative "episode-centric" readings rather than unrelated topic assignments. 51
8. **Topic differences (GB)** 52
9. **Topic differences (HR)** 53
10. **Topic differences (AT)** 54
11. **Macroeconomics and Health vs. other domains.** Country panels show LIWC-22 z-score differences between the focal domain and all other domains (focal minus others), on the LIWC-22 Test Kitchen z-scale. Warm colors indicate higher values in the focal domain; cool colors indicate lower values. Macroeconomics is distinguished by large positive `money` and lower `moral`/`insight` and `power` relative to other topics, while Health shows lower `politic`/`power`/`money` and slightly higher *Tone*. 57
12. **Coalition vs. opposition** (coalition − opposition). Cells show differences in LIWC-22 z-scores on the LIWC-22 Test Kitchen scale; warm colors indicate higher usage in coalition speech and cool colors higher usage in opposition speech. Across countries, coalitions use more positive tone and more collective language (e.g., `we`), while oppositions score higher on overt political/power vocabulary (`politic`, `power`). 59

13. **Topic–LIWC-22 interaction (Austria).** Heatmap of LIWC-22 z-scores by CAP topic and LIWC dimension. Values are expressed on the LIWC-22 Test Kitchen z-scale and shown as topic-specific deviations around the country's overall level for each LIWC dimension. `politic` and `power` are positive across topics (parliamentary baseline), while differences across columns show domain-specific rhetorical shifts. 60
14. **Topic–LIWC-22 interaction (Croatia).** Positive `politic` and `power` scores appear across domains, while the within-grid differences highlight topic-specific rhetorical emphases beyond this general parliamentary baseline. 61
15. **Topic–LIWC-22 interaction (Great Britain).** The grid shows strong political and power language across domains, with especially pronounced levels in international and security-related topics, and more social/narrative markers in welfare- and citizen-facing domains. 61
16. **Ideological orientation.** Group-mean LIWC z-scores by ideological family, standardised relative to the LIWC-22 Test Kitchen Corpus. Each column is an ideological family, and rows are LIWC-22 dimensions. Left and right blocs share a common "parliamentary" profile (high `politic`, low `i`) but differ modestly on analytic and authenticity markers. These differences are smaller than topic and role effects, which supports the interpretation that institutional position and policy domain matter more for style than simple left–right placement. 63
17. **Gender contrasts** (female minus male). Warm (red) cells denote LIWC-22 categories that women use more than men, cool (blue) cells those that men use more. The overall pattern is one of small and heterogeneous effects: women display slightly higher positive tone and work/certainty talk and slightly lower money/politic scores in several panels, but there is no single, dominant "female style." These differences are weaker and less consistent than topic and role effects. 64

18. **Age-group differences.** Group-mean LIWC z-scores by age band (centred on the overall parliamentary mean), still expressed on the LIWC-22 Test Kitchen z-scale. Columns separate age cohorts and rows show LIWC-22 dimensions. Older MPs talk more in institutional, social, and political terms, whereas younger MPs rely somewhat more on certainty language and money-related vocabulary. First-person singular and authenticity markers are relatively higher among younger speakers. These gradients suggest a life-cycle pattern in which parliamentary language becomes more institutional and less self-focused as careers progress. 65
19. **Austria party tone over time (SPÖ, FPÖ, ÖVP, Grüne).** Three-month moving averages of LIWC-22 Tone by party, with background shading indicating each party's own periods in government (green) and opposition (red). Panel headers report total months in each status. The figure shows clear level shifts when parties move between government and opposition, especially for SPÖ, FPÖ, and Grüne. In contrast, ÖVP spends the entire period in government and exhibits no comparable regime breaks. 67
20. **Political language markers over time** (`politic`, `power`, `moral`, `money`). Each panel plots three-month moving averages for one country. Vertical lines mark major external shocks (e.g., migration crisis, COVID-19, Ukraine war); alternating background bands denote parliamentary terms. `politic` spikes during crises and constitutional moments and slowly mean-reverts, while `money` and `power` follow more cyclical patterns tied to budget and institutional debates. `moral` is comparatively stable. The figure summarises our core temporal finding for political rhetoric: shocks raise the temperature of debate, but the chamber does not remain permanently in "crisis mode." 69
21. **Cognitive-process markers over time** (*Cognitive Processes*, *Insight*, *Cause*, *Certitude*). Three-month moving averages for each country. Cognitive-language scores respond less sharply than overt political markers: they drift upward during periods of intense policy justification (e.g., early pandemic months) and then stabilise. This suggests that parliamentarians adjust how much they explain, justify, and express certainty as the policy environment changes. 70

22. **Temporal focus over time** (*Focus Past*, *Focus Present*, *Focus Future*). Present-focused language dominates throughout, but past- and future-oriented speech move systematically: past focus rises around inquiries and post hoc evaluations, while future focus increases in pre-election and strategy-heavy periods. The figure shows that our pipeline can recover not only which topics are on the agenda but also how far debates look backwards or forwards in time. 71
23. **Economic vs. health topics over time.** Monthly prevalence of *Macroeconomics* and *Health* segments in AT, HR, and GB, smoothed with a three-month moving average. Alternating background bands mark parliamentary terms, and vertical lines indicate major crises such as COVID-19. In all three countries, the onset of the pandemic produces a clear substitution pattern: Health surges while Macroeconomics temporarily recedes, before both series drift back toward pre-crisis baselines. Budget cycles are visible as recurring bumps in Macroeconomics, especially in AT and GB. . 72
24. **Security topics over time.** Prevalence of *International Affairs* and *Defence* segments across AT, HR, and GB (three-month moving averages). Shaded bands mark parliamentary terms; dashed vertical lines highlight key external events such as the 2015–2016 migration crisis and the Russian invasion of Ukraine. International Affairs rises sharply around these shocks and then partially recedes, whereas Defence shows smaller, more episodic increases linked to procurement decisions or force deployments. 73
25. **Cross-lingual embedding consistency.** For each speech we embed both the native text and the ParlaMint English translation and compute cosine similarity between the two vectors. Left panels show the distribution of these similarities: most speeches cluster well above 0.8, indicating that the multilingual encoder treats native and translated texts as near-duplicates in embedding space. Right panels show that similarity increases mildly with length (log tokens), which is expected as longer texts provide more lexical evidence. Together, these diagnostics justify using a single English-based LIWC-22 pipeline for AT and HR with country-wise normalisation. . . . 75
26. **LIWC-22 summary variables over time** (*Analytic*, *Clout*, *Authentic*, *Tone*). Three-month moving averages by country. The series show that parliamentary style is relatively stable in the long run, with temporary deviations around major shocks and government changes. 95

27. **Affect over time** (overall affect, positive tone, negative tone). Positive and negative tone move asymmetrically around crises: negative tone rises sharply in acute phases, while positive tone recovers more slowly. 96
28. **Pronoun use over time** (*I*, *We*, *You*, *They*). First-person singular remains consistently low, while collective `we` trends upward. Direct address `you` fluctuates with institutional routines such as Question Time. 97
29. **Social policy topics over time** (Social Welfare vs. Education). Social Welfare displays sharp spikes around major reform packages and fiscal negotiations, whereas Education shows smoother, longer-term trends. . . . 98

30. **Austria: Analytic (party-level).** Three-month moving averages of *Analytic* by party. Governing parties tend to show slightly higher analytic scores, consistent with their responsibility for explaining and defending policy packages. 99
31. **Austria: Authentic (party-level).** Authenticity fluctuates modestly across parties and time, with no simple alignment to government status. . 100
32. **Austria: moral (party-level).** Moral language shows occasional spikes around debates on social issues and migration but remains relatively stable overall compared to political and power vocabulary. 101
33. **Austria: anger (party-level).** Opposition parties display higher and more volatile anger scores, particularly during contentious legislative periods, while governing parties maintain lower, flatter profiles. 102
34. **Austria: anxiety (party-level).** Anxiety-related language peaks around economic and migration crises and then recedes, with only small differences between parties once status is controlled for. 103
35. **Austria: sadness (party-level).** Sadness increases during national tragedies and commemorative debates, affecting all parties similarly. . . . 104

36. **Croatia: Tone (party-level).** As in Austria, governing periods are associated with warmer language, while opposition periods show lower tone scores, especially during contentious reforms. 106
37. **Croatia: Analytic (party-level).** Analytic style increases in reform-heavy periods and slightly more for governing parties, reflecting their role in presenting complex legislative packages. 107
38. **Croatia: Authentic (party-level).** Authenticity exhibits heterogeneous, party-specific dynamics, with some parties adopting a more personal and self-revealing tone over time. 108

39. **Croatia: moral (party-level).** Moral vocabulary intensifies around debates on family policy, education, and national identity, and is somewhat more pronounced in opposition speech. 109
40. **Croatia: anger (party-level).** Anger spikes in periods of political scandal and contested reforms, particularly within opposition parties, and then returns toward baseline. 110
41. **Croatia: anxiety (party-level).** Anxiety markers rise during economic and institutional crises and affect both government and opposition, though with somewhat higher levels in opposition parties. 111
42. **Croatia: sadness (party-level).** Sadness-related language increases during commemorations, natural disasters, and national mourning periods, cutting across party lines. 112

List of Tables

1. Descriptive statistics of analysed corpora. Each corpus includes plenary sittings with verbatim transcripts and speaker metadata. 25
2. **Example sitting excerpt (GB).** Five consecutive speeches with speaker names, text, ParlaCAP labels, and sequence-aware topic assignments ("Our topic"). The second, third, and fourth utterances are locally about fuel poverty, housing, or environmental concerns, and are therefore labelled as *Social Welfare*, *Housing*, or *Environment* at the utterance level. In context, however, they form part of a broader exchange on the Energy Bill and energy policy. Our segment-first labels follow the episode and assign all five utterances to *Energy*. This illustrates the core evaluation mismatch: context-aware segment classification will systematically disagree with context-free references near agenda boundaries and in generic turns. 27
3. **Share of texts exceeding the 8,192-token window (%).** Values report how often speech texts vs. concatenated segments exceed the encoder window. Speech-level overflow is rare, but 13–23% of segments overflow, motivating the chunk-and-average strategy at the segment level. The last column reports the average number of speeches per segment. 38
4. Selected GMM components (k) by country based on Silhouette score (higher is better) and qualitative topic coherence. 41
5. Topic classification performance. Human tests are label-balanced and length-filtered. Full-corpus scores use ParlaCAP labels. 46
6. Pairwise nominal Krippendorff's α between human annotators and GPT-4o on the ParlaCAP human test sets. 92

1. Introduction

Parliamentary debates are among the most structured forms of political communication. Proceedings are recorded verbatim, speakers are identified by role and party, and discussions follow formal agendas that span multiple speeches across multiple speakers. Despite this rich structure, computational analysis typically treats speeches as independent documents. Classifiers assign topics one utterance at a time, ignoring the fact that agenda items unfold over dozens of speeches and that individual contributions often rely on the surrounding context for interpretation.

This approach has consequences. First, it fragments coherent policy episodes. A budget debate may last two hours, but per-speech classification can produce a dozen label switches when speakers use procedural or generic language or briefly mention something that is not central to the current agenda item. Second, it inflates boundary errors. The transition between agenda items creates ambiguous speeches that make sense in context but confuse isolated classifiers. Third, it limits interpretability. Analysts studying temporal dynamics (e.g., when did COVID-19 dominate the agenda?) or institutional patterns (e.g., does the opposition use more emotional language when talking about economics?) need stable topic labels that reflect what the chamber was discussing, not what each individual utterance happens to mention.

Beyond political science, this structure also matters for **business decision-making and risk management**. Parliamentary debates are a high-frequency public signal of policy priorities, regulatory direction, and the importance of economic or social risks. For investors and companies, these signals shape *policy uncertainty* and expectations about taxes, energy regulation, labor markets, public spending, health systems, and international trade. Even when outcomes are not decided in a single sitting, debate dynamics can foreshadow upcoming legislative changes, coalition stability, or changes in the political feasibility of reforms. A topic and episode-level view therefore complements traditional market and macro indicators: it enables systematic monitoring of *what policy risks are being discussed*, *how intensely*, and *by whom*, which are all inputs to strategic planning, scenario analysis, and investment due diligence. This is consistent with evidence that policy-related uncertainty and geopolitical risk are associated with macroeconomic

and financial outcomes, and that firm-level political risk affects real corporate decisions and commands risk premia (Baker et al. 2016; Caldara et al. 2022; Hassan et al. 2019; Pástor et al. 2013).

We address these limitations with a context-aware pipeline that segments debates into coherent agenda episodes before classification. Using the ParlaMint v5 corpus, we analyse 1.4M speeches from Austria, Great Britain, and Croatia (1996–2022). Our method combines lexical cohesion signals, consecutive-speech similarity, and chairperson agenda cues to detect topic boundaries. We then embed segments, cluster them, and map clusters to CAP domains using keyword extraction and conservative LLM label assignment. In parallel, we compute LIWC-22 psycholinguistic features and aggregate by topic, institutional role, party, demographics, and time.

LIWC-22 is a dictionary-based psycholinguistic instrument that maps each text to a set of summary style variables (e.g. *Analytic*, *Clout*, *Authentic*, *Tone*) and hundreds of more specific categories for cognitive processes (`cogproc`, `insight`, `cause`), social and affective language, and content domains such as `money`, `politic`, `power`, or `moral`. When we later talk about "moral" or "insight" language, we refer to these LIWC-22 dictionaries: they capture the share of tokens in a speech that belong to each category, which we normalise so that positive values indicate above-average usage relative to reference texts, as explained in Section 3.5.

This design prioritises episode-level coherence over per-utterance accuracy. Where a speech uses generic parliamentary language ("I thank the honourable member"), our method inherits the label from its segment. A context-free annotator looking at the speech in isolation would often assign a different label than one that has access to the surrounding debate. Where agenda items blur at boundaries, we accept that isolated utterances may look off-topic even when the segment is correct. Standard evaluations penalise these choices: we obtain macro-F1 of 0.47–0.55 against ParlaCAP automatic labels and 0.43–0.47 on balanced human tests, well below supervised baselines with GPT-4o or human labels (0.66–0.76). But our objective differs. We target applications where debate structure matters: time-series analysis of policy attention, dashboards tracking agenda composition, studies of how institutional roles shape discourse within policy domains, and **risk-monitoring use cases** where policy attention and political rhetoric inform investment decisions, regulatory exposure, and strategic planning.

Our substantive findings validate this approach. Across countries, policy topics and institutional roles systematically shape linguistic style. In Macroeconomic debates we observe more transactional vocabulary, elevated authority markers (Clout), and lower moral/insight language than in other domains. Health debates show lower `power` and

`politic` scores and slightly more positive *Tone*. Coalition speakers use more collective pronouns (`we`), higher *Tone*, and less adversarial language; opposition speakers reverse these patterns. Beyond institutional status, we examine how linguistic style covaries with ideological orientation, gender, and age, and find that these demographic and ideological contrasts are present but noticeably smaller than topic and role effects. These regularities hold across Austria, Croatia, and Great Britain, with interpretable country-specific nuances (e.g., Westminster's more political framing of economic debates).

Temporal analysis reveals crisis-driven substitution. When COVID-19 arrives, Health topic prevalence surges while Macroeconomics falls; both revert toward baseline by 2022. The `politic` marker spikes during acute shocks (migration 2015–16, the Ukraine invasion) and mean-reverts afterward. Party-level traces show that linguistic tone tracks institutional position more than calendar time: in Austria, both SPÖ and FPÖ become warmer in government and cooler in opposition, with shifts aligned to role changes.

1.1. Data

1.1.1. Corpora

We analyse Austria (AT), Great Britain (GB), and Croatia (HR) from ParlaMint v5 (Erjavec et al. 2024; *ParlaMint 5.0 now available* 2025). Each corpus includes transcripts, speaker demographics, parliamentary roles (e.g. government vs. opposition, cabinet vs. backbench), and session metadata (dates, sittings, chambers). We restrict attention to plenary sittings with verbatim transcripts and use both native-language and English machine-translation (MT) versions where available.

1.1.2. Corpora Statistics

The combined sample covers about 1.41 million speeches across 4,627 sitting days.

Table 1.: Descriptive statistics of analysed corpora. Each corpus includes plenary sittings with verbatim transcripts and speaker metadata.

Country	**Period**	**Sitting days**	**Speeches**
Austria (AT)	1996–2022	1,221	231,759
Croatia (HR)	2003–2022	1,708	504,338
Great Britain (GB)	2015–2022	2,209	670,912
Total	—	**4,627**	**1,407,009**

1.1.3. Multilingual Processing

For Austria and Croatia we process both native texts and ParlaMint English translations. This preserves fidelity for within-country analysis (topic modelling and segmentation on native texts) while supporting cross-lingual checks and English-only tools (LIWC-22) on the MT versions (Section 4.5). In practice, we compute embeddings for both language variants and assess how similar they are in embedding space; high similarity suggests that the multilingual encoder treats native and translated texts as near-duplicates.

1.1.4. Reference Labels and Human Annotations

We evaluate against two sources.

(1) ParlaCAP automatic labels. These cover all speeches and are used for full-corpus evaluation. They provide a noisy but large-scale reference for CAP domains at the utterance level.

(2) Human-annotated test sets. Manually verified labels for GB and HR (Kuzman et al. 2025; *ParlaCAP: Comparing agenda settings across parliaments via the ParlaMint dataset* 2025). These tests are label-balanced with 40 instances per CAP code, exclude chairperson speeches, and filter to mid-length speeches. They are also annotated in isolation: human labelers see each speech without the preceding or following context. The balanced design shifts the distribution relative to the full corpus, where a few topics dominate. These tests are therefore context-free at the utterance level. Our model is context-aware and segment-first. Disagreements near agenda boundaries reflect a conceptual difference in what is being annotated—utterance content vs. episode topic—rather than necessarily indicating that one or the other is "wrong."

1.1.5. Dataset Illustration

Table 2 shows five consecutive GB speeches from an energy-policy debate. The topic is stable across turns, but per-speech CAP labels vary: some interventions are coded as Social Welfare, Housing, or Environment because they foreground fuel poverty or local fracking concerns. A segment-based approach instead recognises the broader Energy episode and assigns a consistent topic. Per-speech classification ignores this structure; segmentation preserves it and labels speeches in light of context.

Table 2.: **Example sitting excerpt (GB).** Five consecutive speeches with speaker names, text, ParlaCAP labels, and sequence-aware topic assignments ("Our topic"). The second, third, and fourth utterances are locally about fuel poverty, housing, or environmental concerns, and are therefore labelled as *Social Welfare*, *Housing*, or *Environment* at the utterance level. In context, however, they form part of a broader exchange on the Energy Bill and energy policy. Our segment-first labels follow the episode and assign all five utterances to *Energy*. This illustrates the core evaluation mismatch: context-aware segment classification will systematically disagree with context-free references near agenda boundaries and in generic turns.

Speaker	**Text**	**ParlaCAP**	**Our topic**
Rudd, Amber	Our Energy Bill receives Royal Assent today. It is a vital part of our plan to ensure that our families and businesses have access to secure, clean and affordable energy. We are delivering on our manifesto commitment to end subsidies for onshore wind. We are also using the opportunity to support the Oil and Gas Authority with powers to drive greater collaboration and productivity in the industry. I thank the Bill Committee and my hon. Friend the Minister for making this possible and going through the Bill in such painstaking detail to deliver it.	Energy	Energy
Abrahams, Deborah Angela Elspeth Marie	Evidence from the Universities of Leicester and York has shown that sick and disabled people are particularly at risk of fuel poverty, especially after the recent social security cuts by this Government and the previous coalition. Will the Secretary of State approach the Chancellor again to look at better targeting of warm home discount funding, especially after her rebuff from him just before the Budget?	Social Welfare	Energy
Rudd, Amber	The hon. Lady will be aware that this Government, and this Department specifically, are refocusing our support, as far as possible, on to those who are most vulnerable. We have just closed the consultation on the warm home discount and we are looking at the results. She can rest assured that we will, as far as possible, target it at those who are most in need, which is the right thing to do.	Housing	Energy
Tomlinson-Mynors, Michael James	I have been contacted by a number of constituents who are concerned about fracking in Dorset. What reassurance can the Minister give to me and to my constituents about environmental considerations, about issues of public consultation and letting local residents have their say, and, importantly, about fracking being considered only in appropriate locations?	Environment	Energy
Leadsom, Andrea Jacqueline	I can absolutely assure my hon. Friend that the UK has more than 50 years of safely regulating onshore and offshore oil and gas. We have the best regulatory environment in the world. The Environment Agency looks very carefully at any proposals for hydraulic fracturing, the Health and Safety Executive monitors all activity in that area, and of course local authorities will consult widely with their local communities. I am desperate for local communities to be given the proper facts—that is a really important part of the job for us and for local authorities to do.	Energy	Energy

1.2. Research Questions and Contributions

On this basis, our analysis addresses three research questions:

Research questions.

1. **How does policy topic shape linguistic style?** We show that Macroeconomics and Health have distinct LIWC-22 profiles that are stable across countries and years, suggesting domain-specific communication norms.

2. **Do institutional roles, ideological orientation, or speaker demographics produce systematic stylistic differences?** Coalition vs. opposition contrasts persist after conditioning on topic, and we additionally ask whether gender, age, and left–right placement leave systematic but smaller traces in linguistic style.

3. **How do external shocks affect discourse?** Crises reallocate agenda attention (topic substitution) and temporarily elevate political rhetoric markers, with predictable reversion once acute phases end.

Contributions. We offer three advances:

- **Method:** A segmentation-first pipeline that aligns discovered topics to CAP domains without supervised training on parliamentary labels, scales to millions of speeches, and reduces spurious label switches by respecting agenda structure.

- **Measurement:** Cross-lingual LIWC-22 profiles for 22 policy domains, institutional roles, ideological families, and speaker demographics, based on z-scores that use LIWC-22's Test Kitchen Corpus as a benchmark and support between-country comparison.

- **Evidence:** Consistent stylistic regularities across Austria, Croatia, and Great Britain. Macroeconomics shows stronger money- and authority-related markers together with lower moral/insight language; coalition/opposition speech differs predictably; demographic and ideological contrasts are smaller but detectable; and crises produce interpretable topic and style shifts.

Scope and evaluation philosophy. Our approach deliberately sacrifices per-utterance F1 for interpretability in episode-focused tasks. Balanced human test sets look at speeches in isolation, filter to mid-length speeches, remove chairperson turns, and enforce uniform topic distributions, conditions that amplify the cost of context-aware labelling. Many

speeches in long agenda items use generic language; a context-free annotator may assign a different label than the surrounding debate. Our segment-first design inherits context from adjacent turns, so disagreements with isolated references are expected and do not indicate failure for our objective. The dataset illustration in Section 1.1 demonstrates this in practice.

Outline. The remainder of this thesis is structured as follows. Chapter 2 situates our work in the literature on parliamentary corpora, political communication, topic modelling, and discourse segmentation. Chapter 3 details our segmentation, embedding, clustering, CAP mapping, and LIWC-22 scoring pipeline. Chapter 4 presents evaluation of the topic pipeline, cross-sectional LIWC patterns, temporal dynamics, and cross-lingual embedding diagnostics. Chapters 5 and 6 summarise implications and outline directions for future work.

2. Background and Related Work

2.1. Parliamentary Corpora and Policy Coding

ParlaMint standardises parliamentary proceedings for more than 29 European legislatures. It provides native-language transcripts, English translations, and rich metadata on speakers, parties, and institutional roles (Erjavec et al. 2024; *ParlaMint 5.0 now available* 2025; *ParlaMint: Comparable Parliamentary Corpora* n.d.). This enables cross-national designs while preserving institutional context and has already been used for work on legislative agendas, party competition, and institutional communication.

Other multi-country corpora support comparative research. **ParlSpeech V2** aggregates 6.3M speeches across nine democracies with harmonised metadata (Rauh 2020; Rauh et al. 2020). **ParlLawSpeech** links speeches to bills and laws across eight parliaments, enabling studies of how legislative deliberation translates into enacted law (Schwalbach et al. 2024). The **Hansard Corpus** (1803–2005) supports diachronic study of style and ideology in the GB Parliament (Alexander et al. 2015; Hiltunen et al. 2020). These resources motivate our choice to treat parliamentary speech as comparable and policy-relevant across countries and time, and they illustrate the research demand for methods that scale to multi-decade, multi-million-speech settings.

Policy domains. The Comparative Agendas Project (CAP) taxonomy groups political discussions into 22 major topics with stable definitions and a long tradition in agenda-setting research (Baumgartner et al. 2019). ParlaCAP extends CAP labels to ParlaMint, creating large-scale training and evaluation sets for topic classification at the speech level. While most work uses CAP labels to evaluate classification accuracy or to study topic prevalence over time (Baumgartner et al. 2013; Sebők et al. 2021), we use them as a semantic anchor for unsupervised clusters and as a way to interpret stylistic differences across domains.

2.2. Political Language and Style

Institutions and party incentives shape both what is said and how it is said (Lauderdale et al. 2016; Proksch et al. 2010). Text-as-data methods have become standard tools for measuring positions, issue attention, and messaging strategies (Grimmer et al. 2013). Early work includes Wordfish scaling of party positions from manifestos and parliamentary speeches (Slapin et al. 2008), dictionary-based positioning (Laver et al. 2003), and differential language analysis of framing and valence (Monroe et al. 2008). Studies of parliamentary speech document emotional appeals (Rheault et al. 2016; Rudkowsky et al. 2018), moral framing of policy issues (Del Gennaro et al. 2020), and partisan asymmetries in rhetorical style (Jensen et al. 2012).

Much of this work focuses on either positions (left–right, government–opposition) or on single dimensions such as sentiment. We add a systematic profile of style over time, across policy domains, and across institutional and demographic groups using LIWC-22 (Boyd et al. 2022a). Rather than asking whether individual speeches are more emotional or moralised, we ask how these stylistic dimensions vary across topics (Macroeconomics vs. Health), roles (coalition vs. opposition), party families, gender, and age groups, and how they respond to external shocks.

LIWC-22 and political communication. LIWC-22 was originally developed for psychological and clinical applications, but it has been widely adopted for naturalistic text, including political speeches, debates, and social media (Tausczik et al. 2010). Core references cover instrument design and psychometrics (Boyd et al. 2022b; Pennebaker et al. 2015), while more recent work evaluates its validity in diverse corpora and languages (McDonnell et al. 2020). In political communication, LIWC-based studies have examined, among other things, how leaders balance emotional and analytic registers, how moral and power language correlate with leadership evaluations, and how parties differ in their use of collective pronouns (Decter-Frain et al. 2016; Jordan et al. 2019; Körner et al. 2022; Sylwester et al. 2015). We build on this tradition but combine LIWC with a sequence-aware topic pipeline: instead of treating LIWC scores in isolation, we condition them on the policy domain, institutional role, ideological family, gender, and age of the speaker.

Beyond coalition vs. opposition, Government–opposition contrasts are a natural starting point because they encode clear differences in incentives (governing vs. scrutinizing). However, a rich body of work in political sociology and legislative studies suggests that gender, age, and ideology also matter for how politicians speak and the rhetorical strategies they use (Bäck et al. 2019; Hanretty et al. 2025; Hirst et al. 2014).

We therefore situate our coalition–opposition analysis within a broader design that also asks whether left vs. right parties, women vs. men, and younger vs. older MPs differ in their use of analytic language, moral appeals, collective pronouns, or political vocabulary. Our results suggest that these additional contrasts are detectable, but smaller, once we account for topic and role.

2.3. Political Discourse as a Signal for Markets and Risk Management

Parliamentary speech is not only a record of democratic deliberation; it is also a public signal that can affect expectations in the economy. For businesses, policy changes alter costs, constraints, and opportunities through taxation, regulation, procurement, labour rules, environmental standards, and trade policy. For financial markets, the same information can influence risk premia by shifting expectations about inflation, growth, sectoral support, or geopolitical exposure. This makes political text a plausible input to **political-risk monitoring**, **scenario planning**, and **investment research**, especially when combined with quantitative indicators.

A large literature links policy-related uncertainty to macroeconomic and financial outcomes. A widely used benchmark is the Economic Policy Uncertainty (EPU) index, which quantifies uncertainty based on news coverage and related signals and is associated with variation in investment and broader economic activity (Baker et al. 2016). At a more micro level, firm-specific exposure to political risk can be measured from corporate disclosures and is associated with real decisions such as investment and employment, underscoring that political risk is not only a country-level phenomenon but also an unevenly distributed firm-level exposure (Hassan et al. 2019). Related asset-pricing work shows that political uncertainty can command risk premia and affect expected returns, providing a direct channel through which political developments become financially salient (Pástor et al. 2013). Finally, geopolitical shocks represent another major risk channel; the news-based Geopolitical Risk (GPR) index documents sharp spikes around adverse events and links these fluctuations to macro-financial dynamics (Caldara et al. 2022).

A practical challenge is that decision-relevant signals in debates are often *episode-level* rather than speech-level. Long agenda items accumulate information gradually: early speeches introduce motivations and constraints, later speeches reveal coalition fault lines or amendments, and closing statements signal whether a measure is likely to pass. Systems that treat each speech independently can therefore produce noisy estimates of what the chamber is "really" focused on, which is precisely the quantity that risk

functions and analysts want to track. In contrast, coherent agenda episodes provide a stable unit that aligns better with real-world tasks such as "monitor energy-policy risk in Q4" or "detect a regime shift in fiscal-policy attention."

Moreover, *how* an issue is discussed can matter alongside *what* is discussed. Changes in linguistic style—for example, rising conflict language, decreasing collective framing, or increasing certainty markers—can indicate changing bargaining dynamics, escalation, or attempts to stabilise expectations. While such signals do not directly predict market movements on their own, they can enrich qualitative interpretation and help prioritise attention in workflows where analysts must triage large volumes of political information.

In this thesis, we connect these business-facing motivations to method design. Our segmentation-first approach is meant to recover the "agenda backbone" of parliamentary sittings, yielding topic runs that are interpretable for monitoring and comparison. The resulting topic series and style profiles can serve as intermediate features for downstream applications such as early warning dashboards, policy-risk trackers, or sector-specific exposure analyses. Even when the goal remains descriptive rather than predictive, framing the pipeline as part of a broader risk-assessment toolkit clarifies why episode-level coherence and interpretability are valued over maximising utterance-level classification scores.

2.4. Topic Modelling with Embeddings

Embedding-based topic models combine dense semantics with interpretable keywords (Grootendorst 2022). The basic idea is to represent each document as a vector in a high-dimensional space, reduce the dimensionality to denoise and reveal structure, and then cluster these representations into topics. Keywords are recovered from the original text using measures such as c-TF–IDF, which highlight terms that are distinctive for each cluster.

For cross-lingual work, multilingual embeddings such as **BGE-M3** (Chen et al. 2024) (used here), **LaBSE** (Feng et al. 2020), **LASER** (Artetxe et al. 2019), or multilingual **SBERT** (Reimers et al. 2020) aim to preserve semantic similarity across languages: translations and near-translations should lie close in embedding space. This is crucial for our setting, where we want to compare debates in German, Croatian, and English and to use English-only tools such as LIWC-22 while preserving cross-lingual comparability.

For reduction and clustering, **UMAP** preserves local and some global structure in lower-dimensional space (McInnes et al. 2018). We use **Gaussian Mixture Models** for clustering (Reynolds 2009). Alternatives such as HDBSCAN can be robust to noise

and variable-density clusters (Campello et al. 2013, 2015), but in our experiments they produced too many outliers for reliable CAP mapping. Topic representations use c-TF–IDF (Grootendorst 2022), with optional extractors such as **YAKE!** (Campos et al. 2020) or **TextRank** (Mihalcea et al. 2004) to identify salient n-grams.

Model selection and internal validation. We select the number of clusters using information criteria (AIC, BIC) (Akaike 1974; Schwarz 1978) and internal indices (Silhouette, Davies–Bouldin, Calinski–Harabasz) (Calinski et al. 1974; Davies et al. 1979; Rousseeuw 1987). Rather than committing to a single metric, we aggregate ranks across metrics (see Chapter 3). This reflects the fact that no single index is universally optimal: we want clusters that are compact and well-separated (internal indices) but not so numerous that they cease to be interpretable or feasible for CAP mapping (information criteria).

In most topic-modelling applications, documents are treated as exchangeable: their order does not matter. This is a poor fit for parliamentary debates, where topics persist across turns and boundaries align with agenda changes. Ignoring order fragments coherent episodes and makes it difficult to reason about when the chamber is "on" or "off" a given policy topic. Our method therefore embeds and clusters segments rather than individual speeches, and we align clusters to CAP domains ex post using LLM-based label assignment. This allows us to retain the semantic benefits of embedding-based topic models while imposing a structured policy taxonomy and respecting debate sequences.

2.5. Sequential Structure in Discourse

Classic segmentation detects topic boundaries from cohesion dips: when the lexical overlap between adjacent blocks of text falls sharply, this is taken as evidence that one topic has ended and another has begun. Hearst (1997) introduced this idea in the TextTiling algorithm, while Choi (2000) proposed rank-based refinements. Topic-aware methods replace raw word overlap with topic similarity, often derived from latent semantic models (Riedl et al. 2012). Bayesian change-point models treat segments as latent units whose parameters are inferred jointly with the text model (Eisenstein et al. 2008). Neural models learn segmentation end-to-end, predicting boundary positions directly from contextualised representations (Koshorek et al. 2018).

Beyond segmentation, sequential topic models relax the assumption that documents are exchangeable. Dynamic Topic Models (Blei et al. 2006), Hidden Topic Markov Models (Gruber et al. 2007), Topics over Time (Wang et al. 2006), and HDP-HMM variants (Fox et al. 2008) explicitly model temporal or sequential dependencies in topic usage. These

models have been applied to news, scientific articles, and legislative text to study how issue attention and framing evolve over time.

Parliamentary debates, however, present a hybrid challenge: they are both sequential and conversational. Topic changes are sometimes announced explicitly by the chair (e.g. "we now move to the next item") and sometimes emerge gradually as MPs introduce and pursue new lines of argument. Previous work on legislative corpora has mostly either ignored this sequential structure (treating speeches as i.i.d. documents) or focused on short-range conversational dynamics (e.g. question–answer pairs) without linking them to policy codings such as CAP (Abercrombie et al. 2020, 2022; Alvarez et al. 2025; Zhang et al. 2017).

We adopt cohesion-based ideas for the parliamentary setting and fuse similarity signals with chairperson cue phrases and metadata. In contrast to fully supervised sentence-level segmentation, our approach is unsupervised but anchored in observable cues: we search for semantic similarity drops that align with chair interventions and enforce minimum segment lengths so that agenda episodes are not fragmented by noise. The result is a set of segments that can be embedded, clustered, and mapped to CAP domains, providing a sequence-aware backbone for both topic and style analysis.

3. Methods

3.1. Pipeline Overview

The workflow has two branches (Figure 1).

Topic modelling. We embed speeches, segment sittings into coherent spans, re-embed segments, reduce and cluster, then assign CAP labels via keywording.

Linguistic scoring. We compute LIWC-22 features at the speech level and aggregate by topic, role, party, demographics, and time.

Both streams meet at aggregation. This allows us to study linguistic style through topic analysis, role differences within topics, demographic and ideological contrasts, and time trends.

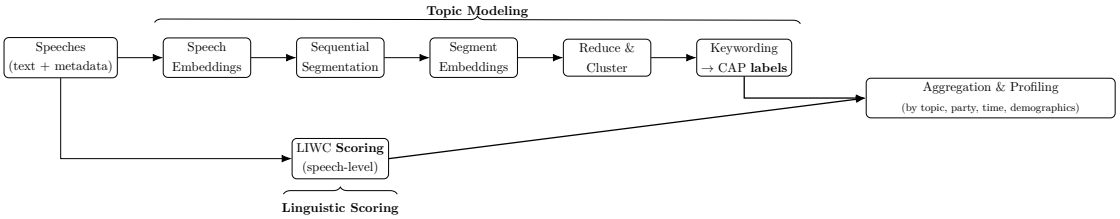


Figure 1.: **Analytical workflow.** Speeches (text + metadata) are embedded, segmented into agenda episodes, re-embedded at the segment level, clustered, and mapped to CAP domains via keywording. In parallel, LIWC-22 is computed per speech and aggregated by topic, role, party, demographics, and time. The key design choice is segment-first topic labelling, which preserves debate context for downstream style analysis.

3.2. Speech and Segment Embeddings

We encode text with `BAAI/bge-m3` (Chen et al. 2024), a multilingual model with an 8,192-token window that outputs **1024-dimensional** embeddings. Longer texts must be truncated or chunked to fit within this window. To avoid truncation, we split very long speeches into overlapping chunks of up to 8,000 tokens with 25% overlap. The 8,000-token limit keeps us safely below the 8,192-token context once model-specific special tokens

(e.g. start/end-of-sequence markers) are added, and the overlap ensures that content near chunk boundaries is represented in more than one vector. We encode each chunk separately and then average the resulting vectors to obtain a single embedding per speech.

For AT and HR we compute embeddings for both native text and English MT to validate cross-lingual consistency. After segmentation (Section 3.3), each segment is re-embedded by concatenating its speeches into a single long string and applying the same chunk-and-average strategy. This captures multi-turn context while keeping the method simple and robust to long agenda episodes.

To quantify how often inputs exceed the model's 8,192-token window, Table 3 reports the share of texts above the limit at both the speech and segment levels: fewer than 0.01% of speeches exceed the window overall, whereas nearly one-fifth of segments do. This motivates applying the chunk-and-average strategy uniformly to both levels when needed.

Table 3.: **Share of texts exceeding the 8,192-token window (%).** Values report how often speech texts vs. concatenated segments exceed the encoder window. Speech-level overflow is rare, but 13–23% of segments overflow, motivating the chunk-and-average strategy at the segment level. The last column reports the average number of speeches per segment.

Country	**Speech (%)**	**Segment (%)**	**Average speeches per segment**
Austria	0.02	20.27	8.10
Croatia	0.00	23.48	13.15
Great Britain	0.00	13.49	20.35
Total	**0.01**	**16.77**	**16.95**

Beyond window usage, the three parliaments also differ in discourse structure. In descriptive statistics (not shown), AT and HR speeches are longer on average and contain fewer very short interjections, whereas GB sittings feature more frequent short turns and interruptions. This is consistent with the higher average number of speeches per segment in GB in Table 3, and it affects both how often segments exceed the token window and how sharply semantic similarity changes at boundaries.

3.3. Sequential Segmentation

Debates follow an agenda. Topics persist across turns and change at item boundaries. Per-speech classification ignores this and adds noise near transitions.

We segment each sitting using two complementary signals combined through an alignment-based approach with automatic parameter optimization.

(1) Semantic similarity drops. We compute cosine similarity between mean embeddings of sliding windows before and after each position. A configurable percentile threshold (default: 95th) identifies the sharpest semantic shifts (Hearst 1997). The window size k is automatically optimized per dataset (details below).

(2) Chairperson agenda cues. Phrases by the chair such as "agenda", "proceed to", "next item", and language-specific variants signal formal topic transitions.

Decision rule. We accept a keyword boundary only if it aligns with a semantic shift (within ± 3 speeches). All remaining semantic boundaries are also accepted. This ensures that detected segments correspond to genuine topic changes rather than arbitrary keyword occurrences.

The alignment requirement filters spurious keyword matches while the semantic signal captures agenda transitions that lack explicit cues. A minimum segment length of 5 speeches prevents over-segmentation.

Automatic window size optimization. We sample 20% of sessions per country and test window sizes from 1 to 10. For each configuration, we evaluate:

- **Keyword alignment score**: Fraction of detected boundaries that align with chairperson cues
- **Semantic quality**: Average of (i) within-segment coherence (mean pairwise cosine similarity) and (ii) between-segment separation (1 minus mean cross-segment similarity)

The final score combines both components with equal weight (50%–50%). We select the window size that maximizes this composite score. This data-driven approach adapts to country-specific discourse patterns without manual tuning.

Implication for GB performance. British parliamentary data contain fewer explicit agenda cues from the chair and more short turns. This weakens the keyword signal and makes semantic windows noisier, but is partly compensated by stronger reliance on semantic detection. We discuss the consequences for evaluation in Section 4.1.

Keywords used. *English*: `agenda`, `proceed`, `point`, `item`, `topic`, `next`, `following`, `move on`.

German (AT): `tagesordnung`, `tagesordnungspunkt`, `punkt`, `verhandlung`, `behandlung`, `nächster`, `weiter`, `fortsetzen`.
Croatian (HR): `dnevni`, `red`, `točka`, `tačka`, `sljedeći`, `sljedeće`, `prijedlog`, `zakon`, `tema`, `nastavljamo`, `prelazimo`.

3.4. Clustering and CAP Alignment

Reduction and clustering. We reduce segment embeddings with UMAP (McInnes et al. 2018) and cluster with a Gaussian Mixture Model (GMM). K-means over-balanced cluster sizes and forced spherical structure, while HDBSCAN produced too many outliers for reliable CAP mapping at our scale.

Why UMAP? UMAP preserves local neighbourhoods under a cosine metric, handles non-linear structure, and acts as a strong denoiser before clustering. We set $n_{\text{neighbors}} = 15$, $n_{\text{components}} = 10$, min_dist $= 0.05$, metric=cosine.

Dimensionality reduction (why 1024 $\rightarrow$ 10). Clustering directly in 1024D is brittle: in high dimensions pairwise distances concentrate (the "curse of dimensionality"), making separation unreliable because almost everything is similarly "far apart." We therefore reduce embeddings with UMAP to **10 dimensions**, which (i) denoises, (ii) preserves neighbourhood structure under a cosine metric, and (iii) yields more stable, well-separated clusters for downstream modelling.

Why GMM (vs. k-means/HDBSCAN)? GMM models *elliptical* cluster shapes and yields *soft* assignments (posteriors), which are useful for downstream aggregation and for identifying mixed or boundary segments. K-means assumes equal-variance spheres and was sensitive to scale, yielding clusters that were too similar in size where our data structure suggests that they are not; HDBSCAN's outlier sensitivity fragmented coherent policy areas in our setting, and the outlier percentage was too high even with strict constraints.

k selection for GMM. We select the number of mixture components k by scanning a grid $k \in [100, 300]$ in steps of 5 and computing the Silhouette score for each fitted model. For each country we choose the k that maximises the Silhouette score, subject to a basic interpretability check (excluding choices that produce extremely small or extremely large clusters). This keeps the selection procedure simple and transparent while still favouring compact, well-separated clusters.

Table 4.: Selected GMM components (k) by country based on Silhouette score (higher is better) and qualitative topic coherence.

Country	**Selected k**
Austria (AT)	180
Great Britain (GB)	170
Croatia (HR)	185

Topic representations (c-TF–IDF). We compute class-based TF–IDF on segments per GMM component. We keep unigrams and bigrams, **drop terms that occur in more than 95% of segments**, and **drop very rare terms** (document frequency below **0.5%** of segments or `min_df=5`, whichever is higher). We cap the vocabulary per country to control memory, and extract the top **15** n-grams per component. Ten UMAP dimensions were chosen after visual inspection of cluster separation and topic coherence.

Many-to-one CAP mapping via LLM. Each component (subtopic) receives a single CAP domain in two steps:

(1) Keywording. Use the c-TF–IDF n-grams to summarise the component.

(2) LLM label assignment. Feed the **15** n-grams (unigrams/bigrams) plus country/language context into a conservative classifier prompt (see Appendix B). The **output is exactly one CAP category**. We set temperature=0 for determinism and default to *Other* or *Mix* when uncertain. In total we ran 535 assignments (180 + 170 + 185). This procedure preserves the unsupervised nature of clustering while anchoring topics in a widely used policy taxonomy.

3.5. Psycholinguistic Profiling

We score each speech with LIWC-22 (Boyd et al. 2022a). The instrument covers summary variables, grammar, psychological processes, and content domains. For AT and HR we score English MT texts; following the LIWC authors' guidance, it is preferable to translate to English and then apply LIWC-22 than to use non-English dictionaries ad hoc.

LIWC-22 ships with a large reference dataset, the *Test Kitchen Corpus*, which summarises how often each category appears in many genres of everyday English text (Boyd et al. 2022a). For each LIWC variable c the manual reports a mean μ_c and standard deviation σ_c on this corpus. We convert our raw percentages p_c (share of tokens in

category c) into *z-scores*

$$z_c = \frac{p_c - \mu_c}{\sigma_c}.$$

A value of $z_c = 0.3$ for `moral`, for example, means that the text uses moral words about 0.3 standard deviations more often than the texts in the Test Kitchen Corpus. In many figures we look at *differences* in these z-scores between groups (e.g. Macroeconomics minus "other topics", coalition minus opposition), while in others we plot raw group means on this same z-scale.

In the visualisations below we use two closely related types of values:

- Some heatmaps show *raw* LIWC z-scores relative to the Test Kitchen Corpus (for example Figures 2, **??**, 16, 17, and 18). In these panels entire rows such as `politic` or `power` can be positive for every group. This does *not* mean that all groups are equally political; it simply reflects that parliamentary language, as a whole, uses more political and power-related vocabulary than the general texts in the LIWC reference corpus.

- Other heatmaps show *differences* between two groups on the same LIWC scale (for example Figures 11 and 12). There we plot z-score differences such as "Macroeconomics minus all other topics" or "coalition minus opposition". A cell value of +0.3 in these figures means that the focal group uses words from that dictionary *0.3 standard deviations more often* than the comparison group, measured on the same LIWC reference scale.

Throughout the results we focus on the size and sign of these effects, and on whether patterns are stable across countries, rather than on formal significance tests.

Because we work with more than 1.4M speeches and many millions of tokens, even very small shifts away from zero are statistically significant in the usual sense. We therefore focus on how *large* differences are (effect size and direction) and on whether patterns are stable across countries and over time, rather than on p-values.

What LIWC-22 measures (dictionary-based). LIWC-22 is a *dictionary approach*: each category is a curated list of words/stems. A score for a text is the *percentage of tokens* in that category's dictionary (tokens in category divided by total tokens). Because scores are proportions, they are comparable across texts after standardisation by the Test Kitchen norms. Summary variables such as *Analytic*, *Clout*, *Authentic*, and *Tone* are derived from combinations of dictionary counts and internal scaling; content categories

such as `money`, `politic`, `power`, `moral`, or `insight` are direct percentages that we then express as z-scores.

Feature coverage and overview figure. Figure 2 shows a heatmap for the main dimensions we use as a benchmark: parliamentary language in each country relative to general-population norms from the LIWC-22 Test Kitchen Corpus.

Visual conventions. All following LIWC-22 heatmaps follow the same rule.

- **Top five rows fixed:** *Analytic*, *Clout*, *Authentic*, *Tone*, *power*.
- **Next five rows:** the five largest absolute differences for the contrast of interest.
- **Country sample sizes** appear below panels when relevant.

3.6. Code and Reproducibility

All preprocessing scripts, segmentation and topic-modelling pipelines, and notebooks used to generate the results and figures in this thesis are available in a public GitHub repository: https://github.com/pavlesav/master-thesis. The repository includes code for data preparation, embedding and clustering, CAP mapping, and LIWC-22 analyses, together with configuration files and documentation to support reproducibility.

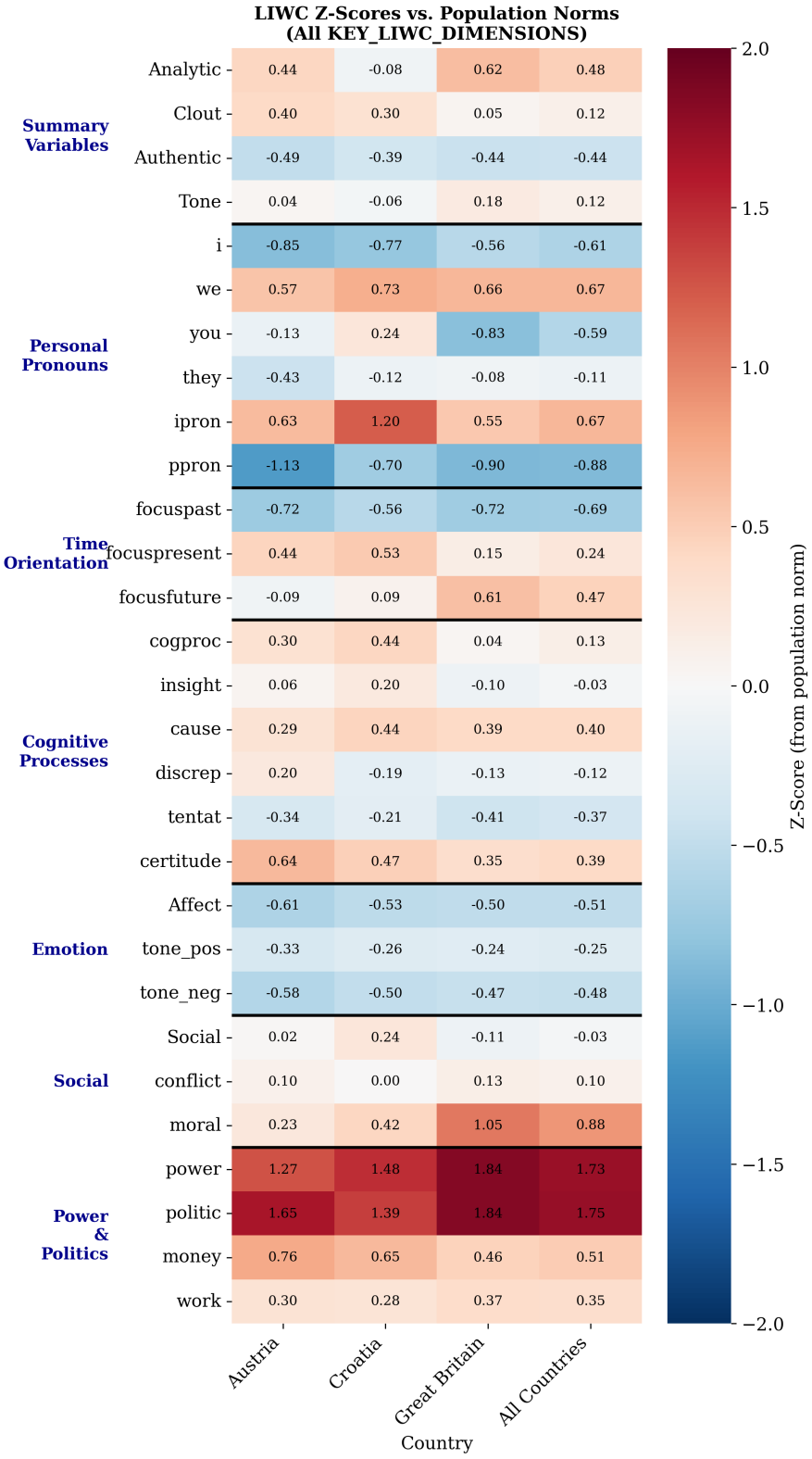


Figure 2.: **Political discourse vs. population norms.** Country-wise LIWC-22 z-scores for parliamentary speech relative to general-population benchmarks from the LIWC-22 Test Kitchen Corpus. Positive cells indicate categories that are over-represented in parliamentary language (e.g., political and power vocabulary), while negative cells indicate under-use. This overview panel provides the reference baseline for all later figures: topic-, role-, ideology-, demographic-, and time-specific effects should be interpreted as deviations from the already distinctive style of parliaments shown here.

4. Results

We first evaluate how the sequence-aware topic pipeline aligns with existing CAP-style labels and human annotations, including confusion matrices and shifts in topic distributions. We then turn to LIWC-22-based stylistic profiles across policy domains and institutional roles, followed by ideological, gender, and age contrasts. Next we analyse temporal dynamics of agenda attention and political rhetoric, before closing with cross-lingual embedding diagnostics that justify the English-based LIWC-22 pipeline for Austria and Croatia.

4.1. Evaluation of Topic Modelling

4.1.1. Task mismatch and evaluation philosophy

Our model anchors labels to local agenda context, whereas both ParlaCAP automatic labels and the human test sets evaluate isolated utterances. In long episodes, a generic turn can look off-topic when taken alone. Our labels follow the segment: if a short procedural question occurs in the middle of an immigration debate, we label it as *Immigration* rather than as *Government Operations*. The example in Section 1.1 (Table 2) shows this behaviour explicitly.

From an evaluation standpoint, this introduces a systematic mismatch. Context-free annotators and classifiers try to infer the topic from the local text alone; context-aware systems like ours use the surrounding episode as an additional signal. Near agenda boundaries, under heavy use of generic parliamentary language, these two perspectives will disagree even when both are plausible. Per-utterance F1 therefore underestimates performance for our actual use case: reliable, interpretable labelling of *episodes* rather than individual speeches.

4.1.2. Scores against reference labels

Table 5 summarises alignment with reference labels at two levels.

Table 5.: Topic classification performance. Human tests are label-balanced and length-filtered. Full-corpus scores use ParlaCAP labels.

	vs. ParlaCAP (full corpora)		vs. human tests
Country	**Macro-F1**	**Micro-F1**	**Macro-F1**
Austria (AT)	0.55	0.57	—
Great Britain (GB)	0.47	0.51	0.43
Croatia (HR)	0.52	0.56	0.47

Reported supervised baselines on human tests (from ParlaCAP): GPT-4o macro-F1 0.761 (GB), 0.657 (HR)

Against ParlaCAP labels on the full corpora we obtain macro-F1 between 0.47 and 0.55 and micro-F1 between 0.51 and 0.57. On balanced human tests (GB and HR) macro-F1 is lower, which reflects both distribution shift and our design choice to prioritise sequence structure. Supervised systems tuned on these references perform markedly better at the utterance level (Kuzman et al. 2025; *ParlaCAP: Comparing agenda settings across parliaments via the ParlaMint dataset* 2025), but they do not directly address the episode-level coherence objective that motivates our pipeline.

4.1.3. Human agreement as a ceiling

The ParlaMint/ParlaCAP team reports moderate agreement among human coders (Krippendorff's $\alpha \approx 0.59$–0.68) with GPT-4o at a similar level (Appendix A). This matters for interpretation: even ideal systems cannot exceed human disagreement by a large margin on this labelling scheme. Our unsupervised, sequence-aware pipeline therefore operates under a double handicap: it is evaluated (i) at a finer granularity (utterances) than it was designed for (segments) and (ii) against references with inherent noise and ambiguity in boundary regions.

4.1.4. Error structure and confusion patterns

Figures 3 to 7 summarise where our segment-informed labels disagree with (i) ParlaCAP and (ii) human test annotations (GB/HR only). Across countries, off-diagonal mass is not random: it clusters in substantively adjacent CAP domains, such as Macroeconomics vs. Domestic Commerce, International Affairs vs. Foreign Trade, and Social Welfare vs. Health. In qualitative inspections, many of these cases correspond to genuinely mixed or boundary episodes (e.g., economic implications of foreign policy).

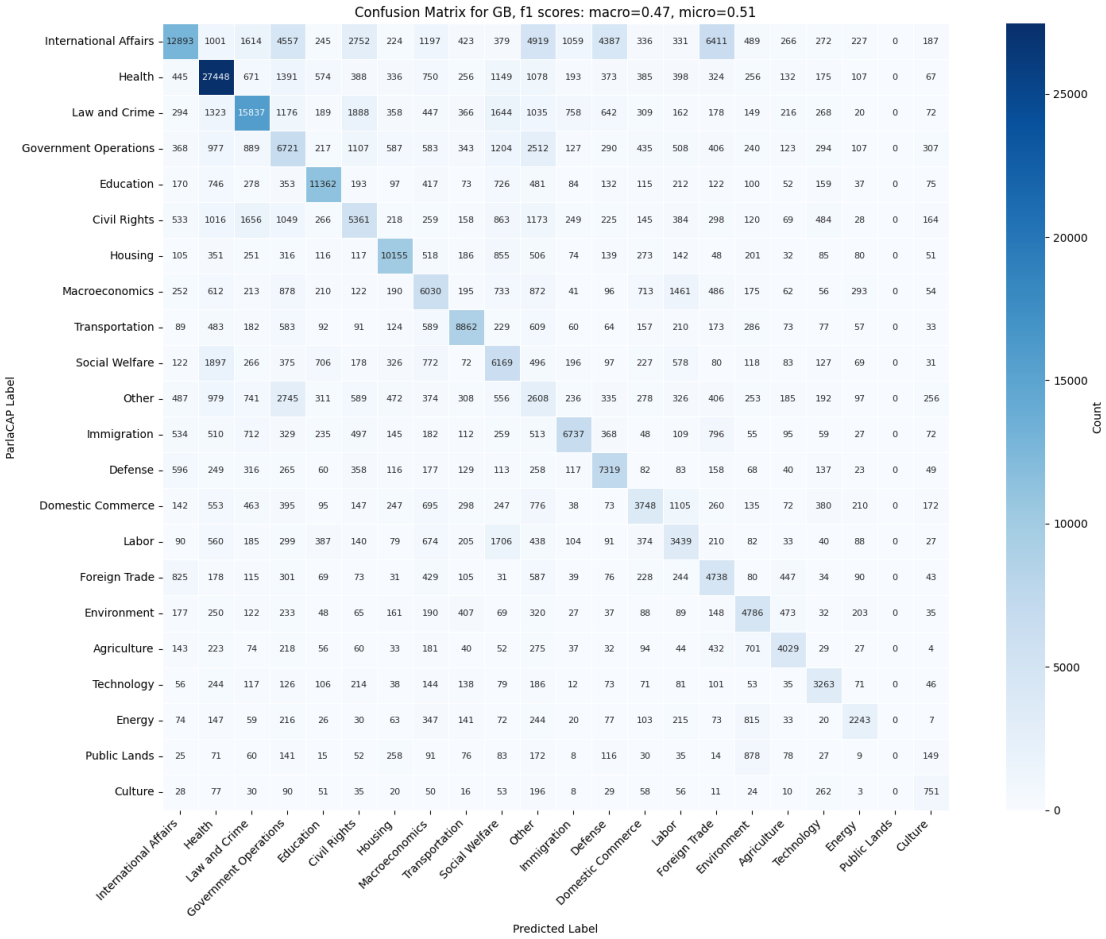


Figure 3.: **Great Britain (GB): model vs. ParlaCAP.** Confusion matrix of predicted CAP domains (x-axis) against ParlaCAP labels (y-axis) for the full corpus; the diagonal indicates agreement. Off-diagonal mass concentrates in conceptually adjacent domains (e.g., welfare–health, macro–commerce), consistent with boundary and mixed-episode cases.

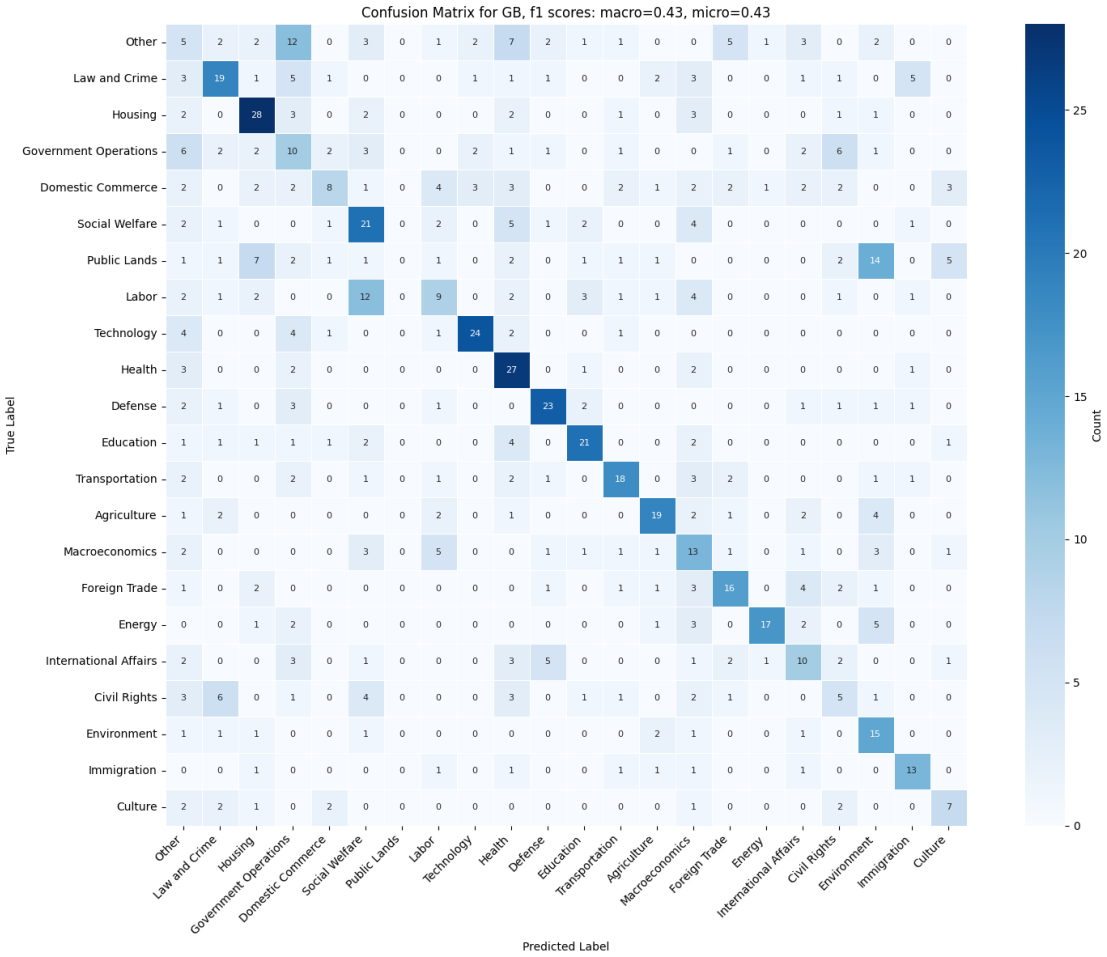


Figure 4.: **Great Britain (GB): model vs. human test labels.** Compared to ParlaCAP, human labels typically reduce ambiguity in boundary speeches, but the same broad structure remains: most off-diagonal mass lies between neighbouring policy areas. This supports the interpretation that many "errors" are boundary disagreements (what the episode is *mainly* about), not failures to identify the general policy region.

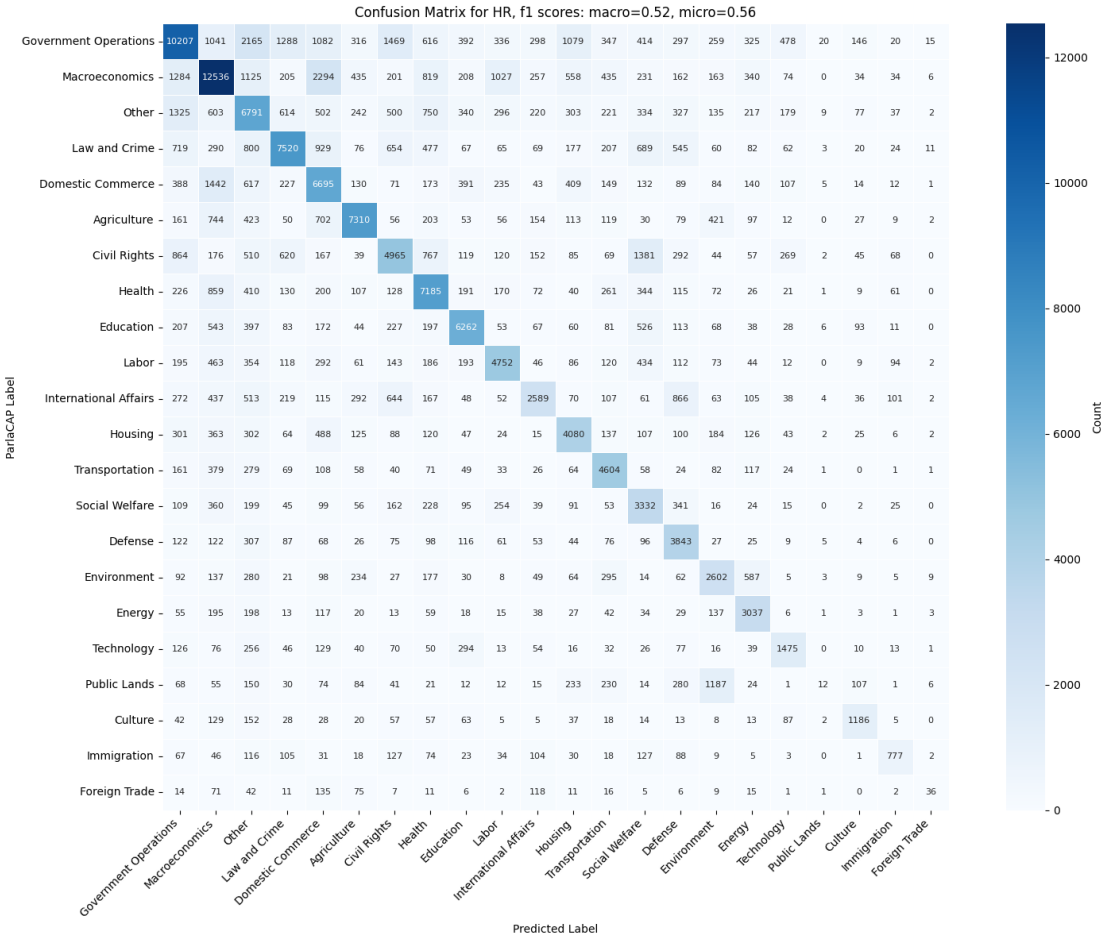


Figure 5.: **Croatia (HR): model vs. ParlaCAP.** HR shows the same pattern as GB: misclassifications are concentrated between related domains. In practice, these often correspond to agenda items that naturally mix policy frames (e.g., public administration reforms discussed alongside fiscal implications), where segment-level labels prioritise episode coherence.

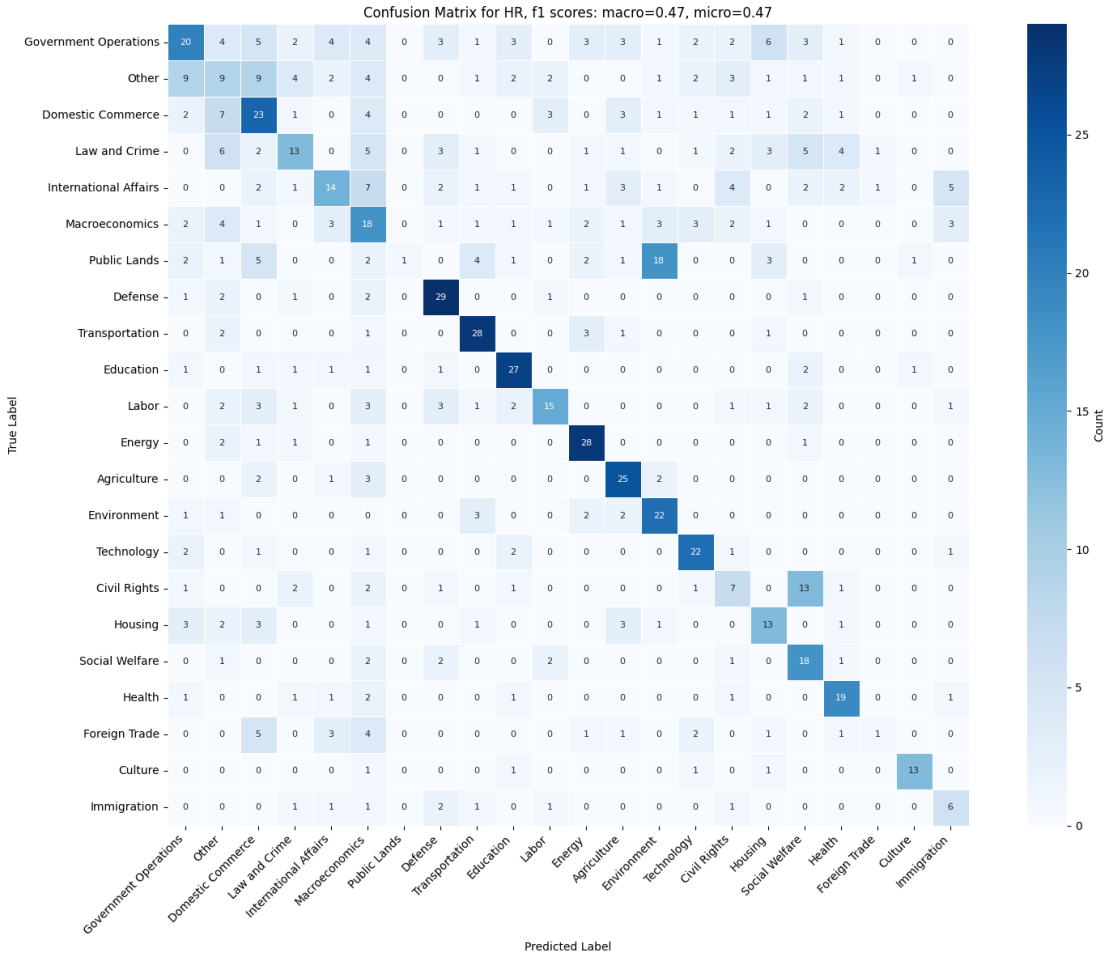


Figure 6.: **Croatia (HR): model vs. human test labels.** Agreement improves relative to ParlaCAP, but remaining disagreements again cluster around adjacent categories. This suggests that our pipeline preserves the correct policy neighbourhood even when it differs on fine-grained boundaries, which is the expected trade-off when enforcing segment-level consistency.

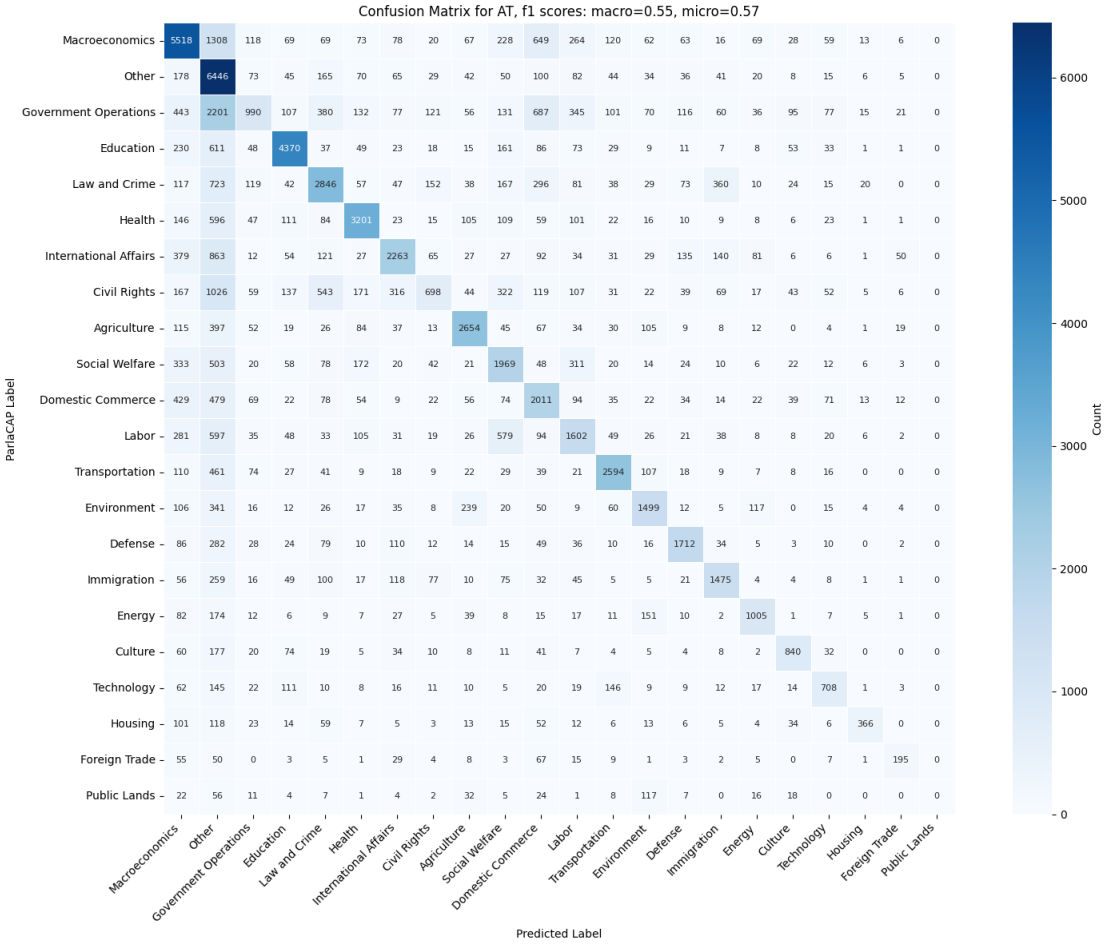


Figure 7.: **Austria (AT): model vs. ParlaCAP.** Austria has no human test set in our evaluation, so we report agreement against ParlaCAP on the full corpus. The same qualitative structure holds: disagreements are dominated by adjacent or thematically overlapping CAP domains. This is consistent with the interpretation that differences often reflect alternative "episode-centric" readings rather than unrelated topic assignments.

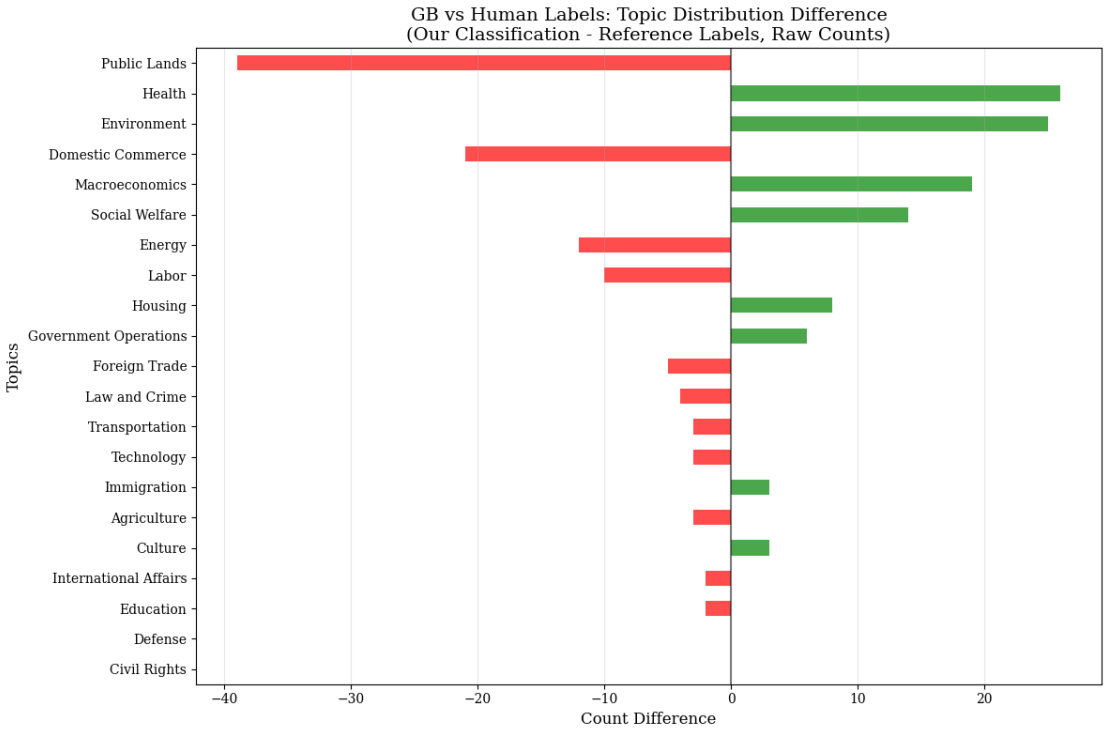


Figure 8.: **Topic distribution differences (GB vs. human labels).** Bars show the difference in topic frequency between our segment-informed labels and the human reference labels (our minus reference; raw counts). Positive values indicate domains assigned more often by our pipeline, negative values more often by the human labels. Differences are concentrated in a subset of categories, reflecting systematic shifts between episode-level and utterance-level labelling.

To complement the confusion matrices, Figures 8 to 10 compares how our labels change the topic distribution relative to the reference labels. For GB and HR we use the human test sets; for Austria we compare against ParlaCAP. Positive bars indicate topics that our model assigns more often than the reference; negative bars mark topics where the reference labels are more common.

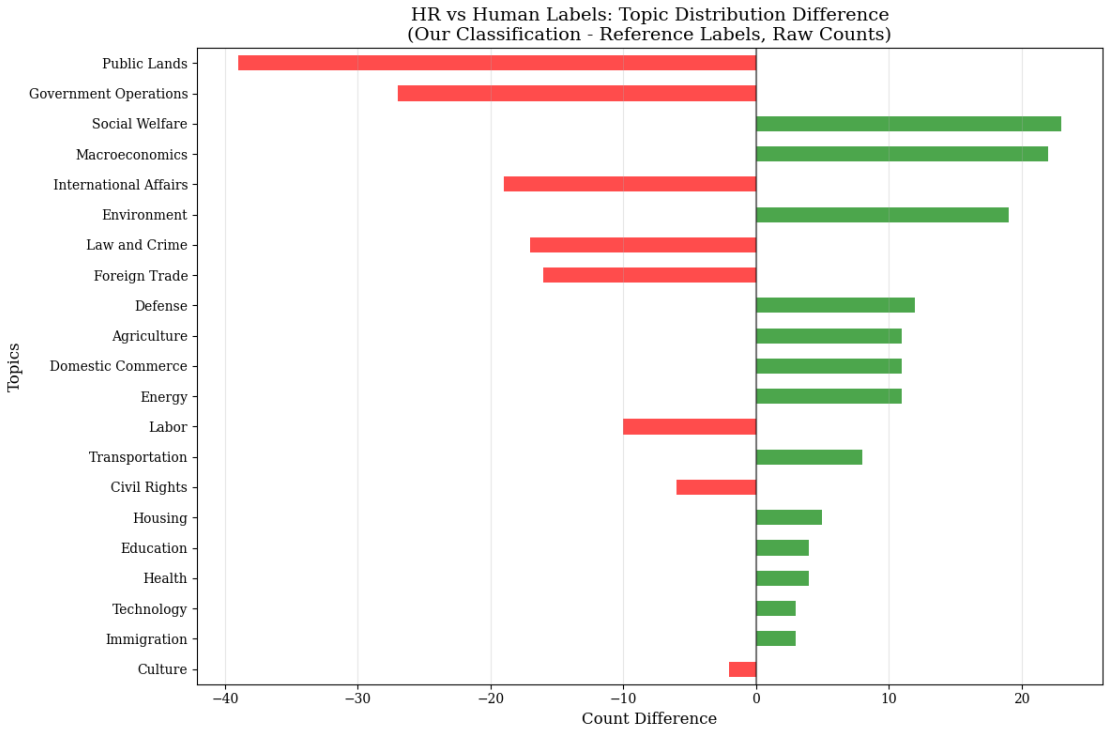


Figure 9.: **Topic distribution differences (HR vs. human labels).** Bars show how many more (or fewer) segments we assign to each CAP domain compared to the human test labels (our minus reference, raw counts). As in GB, the strongest differences concentrate in thematically related domains, consistent with the confusion-matrix structure. This suggests that disagreements are often about *where to place the boundary* between adjacent topics (e.g., macroeconomic framing inside a broader social-policy debate), rather than confusion between distant areas. Overall, the pattern supports the interpretation that our labels preserve the correct policy "neighbourhood" while enforcing episode coherence across surrounding speeches.

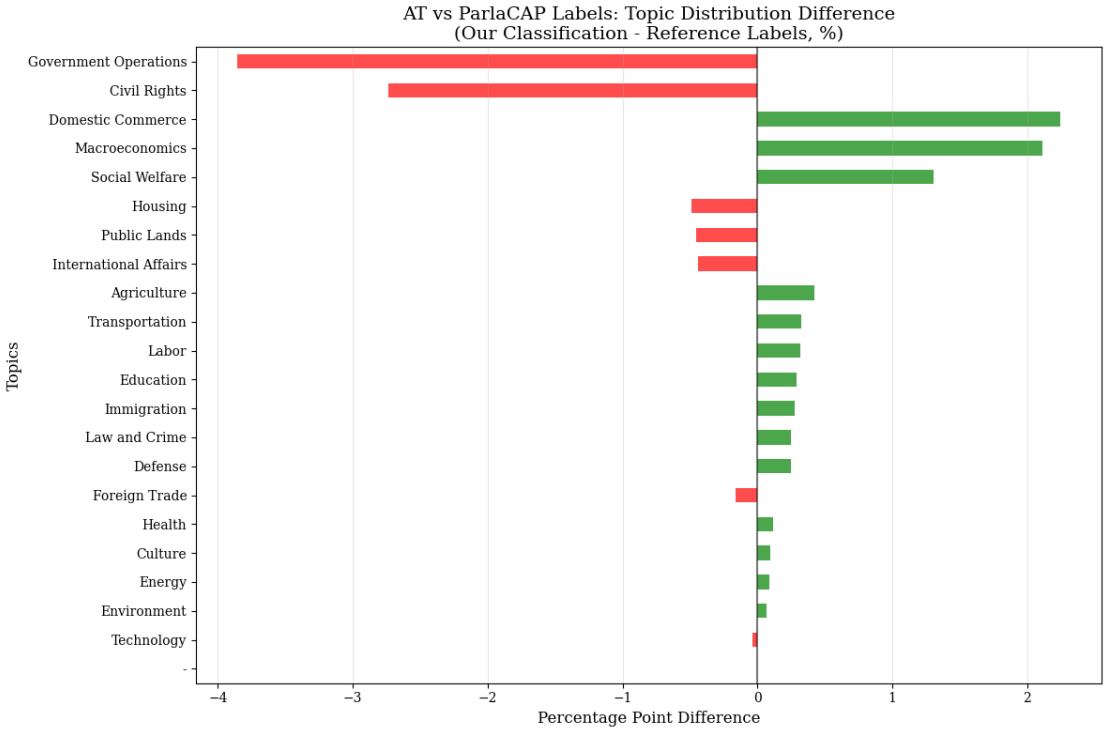


Figure 10.: **Topic distribution differences (AT vs. ParlaCAP).** Austria has no human test set in our evaluation, so we compare our labels against ParlaCAP on the full corpus. Unlike GB and HR, this plot shows differences in *percentage points*, because the full-corpus counts are much larger and raw differences would dominate the figure. Positive bars indicate CAP domains that our segment-informed pipeline assigns more frequently than ParlaCAP; negative bars indicate domains that ParlaCAP assigns more often. Larger shifts typically appear in categories that are either broad (where surrounding context strongly influences the episode label) or rare/ambiguous (where short interventions can be pulled into the dominant agenda topic). This comparison should therefore be interpreted as *distributional drift* between two labelling schemes, not as a direct measure of correctness.

4.1.5. Country-specific challenges

The British data set is particularly challenging for our segmentation procedure: it contains fewer explicit agenda cues from the chairperson, and long sequences of short interventions make semantic boundaries fuzzier. In internal diagnostics (not shown), fewer chair-phrase hits align with strong embedding similarity drops, which weakens one leg of our decision rule. This likely contributes to lower macro-F1 in GB relative to AT and HR. At the same time, GB displays some of the clearest temporal and stylistic patterns in our analysis (e.g., higher politicisation of Macroeconomics, a strong coalition–opposition politic gap), suggesting that the episode-level structure is still useful even when utterance-level agreement is modest.

4.1.6. What these scores mean for users

For users interested in per-speech classification - for example, building training sets for supervised models or coding short interventions in isolation—our pipeline is not the best tool; recent supervised transformers and GPT-4o-based labellers are more appropriate (Kuzman et al. 2025). For users interested in *agenda episodes*, temporal dynamics, or role-based stylistic patterns, our sequence-aware labels offer three advantages:

1. Fewer spurious topic switches within long items;
2. Easier interpretation of time-series plots and dashboards, because contiguous runs correspond to single CAP domains;
3. A natural unit of analysis (segment) that aligns better with how MPs, journalists, and observers describe "what the house was discussing."

The rest of the Results chapter is written with these users in mind.

4.2. Linguistic Profiles Across Policy Domains and Roles

4.2.1. Macroeconomics and Health

Policy domains have distinctive linguistic profiles. Figure 11 contrasts *Macroeconomics* and *Health* with all other topics using country-wise z-score differences (target topic minus "other topics"). Values near zero indicate no stylistic shift; positive/negative values indicate over-/under-use within the focal topic.

Macroeconomics (money and authority markers). Across AT, HR, and GB, Macroeconomics is consistently more *money*-driven (around $+1.4\,z$ on average) and shows

a small, uniform increase in *Clout* ($\sim +0.06$–$+0.11$). The high `money` scores reflect the obvious centrality of fiscal and financial vocabulary, but the cross-country stability is notable given institutional differences between the Austrian, Croatian, and Westminster systems. At the same time we observe lower use of reflective or moralising language: `insight` and broader `cogproc` are modestly lower (≈ -0.1 to -0.3), and `moral` is lower in every country. `power` is also depressed (≈ -0.3 to -0.45), suggesting that economic debates rely less on overt authority and more on detailed financial argument. *Analytic* is slightly negative or near zero; the largest differences therefore appear in domain-specific dictionaries (`money`, `moral`, `insight`, `power`) rather than in the summary variable.

The GB panel introduces a country-specific nuance: a mild uptick in `politic` ($\sim +0.30$) relative to other topics, indicating that Westminster economic debates are linguistically more overtly political than their Austrian and Croatian counterparts. This aligns with the stronger two-party conflict structure in the UK, where macroeconomic questions are central to partisan competition.

Health (lower politic/power markers). Health systematically shows lower political conflict markers: strong negatives on `politic` (≈ -0.29 to -0.65) and `power` (≈ -0.28 to -0.58), together with lower `money` (≈ -0.26 to -0.38). *Tone* is slightly higher in most panels ($\sim +0.05$ to $+0.11$), consistent with a more positive or reassuring register, especially during discussions of service provision or public-health achievements. *Analytic*, *Authentic*, and *Clout* hover near zero or slightly positive, and `conflict` is flat to slightly negative, except for a small positive in HR which may reflect contentious reforms or hospital funding debates. Relative to other domains, health debates therefore contain fewer terms from the `politic`, `power`, and `money` dictionaries while remaining similar on most other dimensions.

Takeaway. Compared to other domains, Macroeconomics shows higher `money` and slightly higher *Clout* together with lower `moral`, `insight`, and `power` scores, while Health shows lower `politic`, `power`, and `money` and slightly higher *Tone*. These regularities hold across countries, with a GB-specific nuance of slightly more political framing (`politic`↑) in economic debates. The contrast illustrates the broader claim that policy domains come with characteristic rhetorical styles, not just different vocabularies.

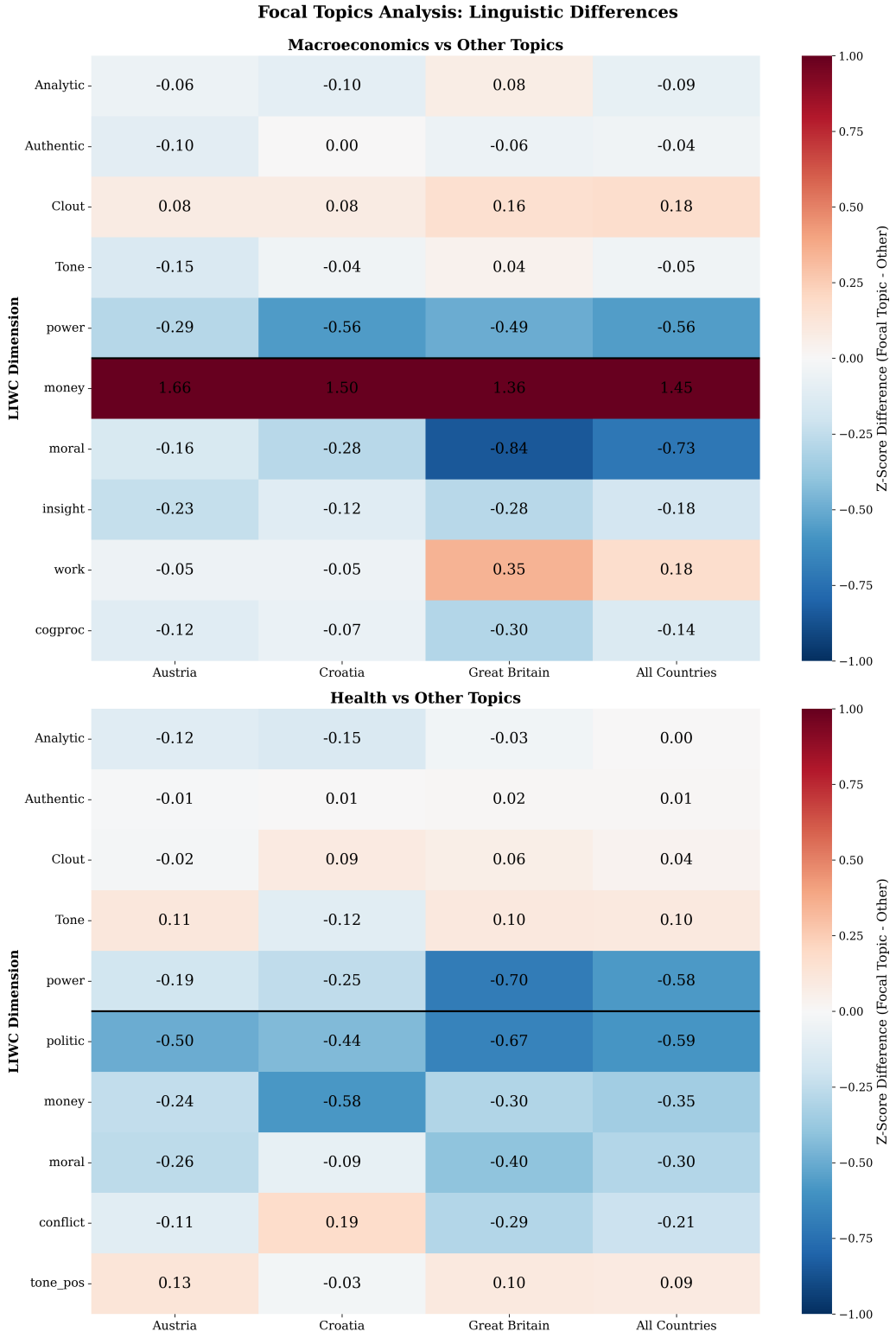


Figure 11.: **Macroeconomics and Health vs. other domains.** Country panels show LIWC-22 z-score differences between the focal domain and all other domains (focal minus others), on the LIWC-22 Test Kitchen z-scale. Warm colors indicate higher values in the focal domain; cool colors indicate lower values. Macroeconomics is distinguished by large positive `money` and lower `moral`/`insight` and `power` relative to other topics, while Health shows lower `politic`/`power`/`money` and slightly higher *Tone*.

4.2.2. Coalition vs. Opposition

Figure 12 reports country-wise z-score *differences* (coalition minus opposition). Positive values indicate features used more by the coalition; negative values indicate features used more by the opposition.

Coalition signature (Tone, pronouns, politic/power). Across all countries, *Tone* is consistently higher for coalition speakers, often by around +0.2 to +0.5 standard deviations. Collective framing rises via `we`, while adversarial address `you` is generally lower or flat. Coalition speech also tones down political contestation markers: `politic` and `power` are negative in every country, with the GB `politic` gap especially large (around −0.75). *Authentic* is slightly higher overall, and *Clout* is roughly neutral, indicating that incumbents do not necessarily project more self-confidence in LIWC terms, but they do soften their style and present themselves as stewards rather than combatants.

Opposition mirror (politic/power and pronouns). Opposition speech shows the mirror pattern: more `politic` and `power`, more direct address (`you`) in AT (and on average), less `we`, and lower *Tone*. These differences are not huge in absolute terms, but they are remarkably stable across three institutional contexts and over long periods of time. Country idiosyncrasies remain small relative to the shared cross-national structure, suggesting that the coalition/opposition divide induces similar rhetorical incentives in all three parliaments.

Takeaway. Coalitions use more collective and positive language (*Tone*↑, `we`↑, `politic`/`power`↓). Oppositions show higher scores on `politic`, `power`, and (in some settings) `you`, and lower *Tone*. These differences are consistent across AT, HR, and GB, with GB exhibiting the strongest politic gap. The figure therefore summarises one of the main empirical contributions of the thesis: institutional role, rather than ideology, is the primary driver of how adversarial or conciliatory parliamentary speech sounds.

4.2.3. Topic–style interactions across all domains

Beyond Macroeconomics and Health, other CAP domains show recognisable stylistic signatures. Security and foreign-policy codes contain higher shares of `politic` and `power` vocabulary; welfare and social-policy codes score higher on social and affective language; and culture-related codes contain more narrative and personal markers. These patterns are easier to see in a full topic–style grid.

Figures 13 to 15 show z-scores by CAP topic and country for a set of key LIWC-22 dimensions. Each cell shows how much a topic deviates from the pooled country–topic mean for that dimension. Several themes emerge:

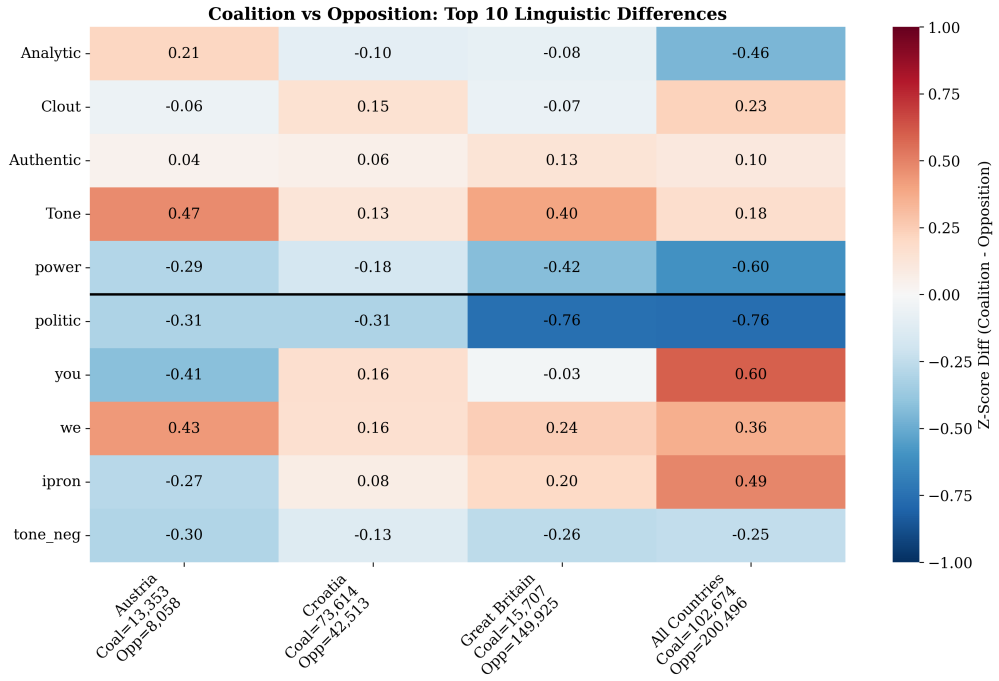


Figure 12.: **Coalition vs. opposition** (coalition − opposition). Cells show differences in LIWC-22 z-scores on the LIWC-22 Test Kitchen scale; warm colors indicate higher usage in coalition speech and cool colors higher usage in opposition speech. Across countries, coalitions use more positive tone and more collective language (e.g., `we`), while oppositions score higher on overt political/power vocabulary (`politic`, `power`).

- **Security and international affairs** are uniformly high on `power` and `politic`, especially in GB, reflecting how central foreign and defence policy is in party competition at Westminster.
- **Law, crime, and immigration** debates combine elevated `politic` and `power` with relatively low collective markers (e.g., pronouns such as `we`), consistent with a focus on control, enforcement, and adversarial framing.
- **Social welfare and health** are comparatively lower on power/politic but higher on social markers and pronouns, matching their focus on citizens' everyday lives.
- **Culture, education, and civil rights** lean more toward *Authentic* language and cognitive-process markers, with somewhat weaker power signals.

These domain-specific signatures are not perfectly sharp—topics blend, and many speeches mix multiple concerns—but the overall structure is robust across countries.

Figures 13 to 15 therefore complement Figure 11 by showing that Macroeconomics and Health are part of a broader, interpretable map of domain-specific rhetorical styles.

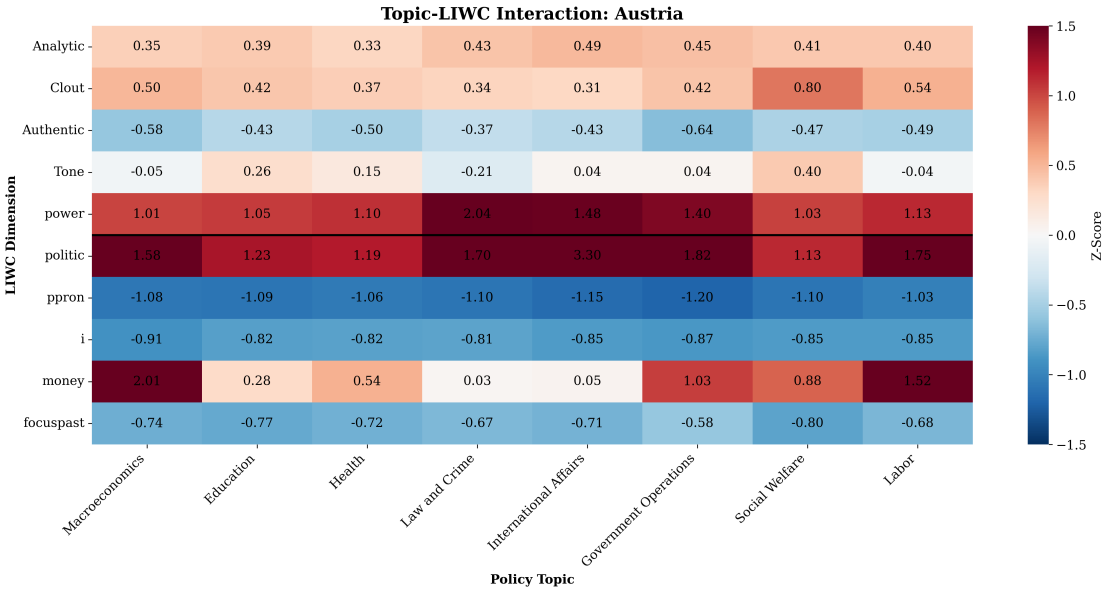


Figure 13.: **Topic–LIWC-22 interaction (Austria).** Heatmap of LIWC-22 z-scores by CAP topic and LIWC dimension. Values are expressed on the LIWC-22 Test Kitchen z-scale and shown as topic-specific deviations around the country's overall level for each LIWC dimension. `politic` and `power` are positive across topics (parliamentary baseline), while differences across columns show domain-specific rhetorical shifts.

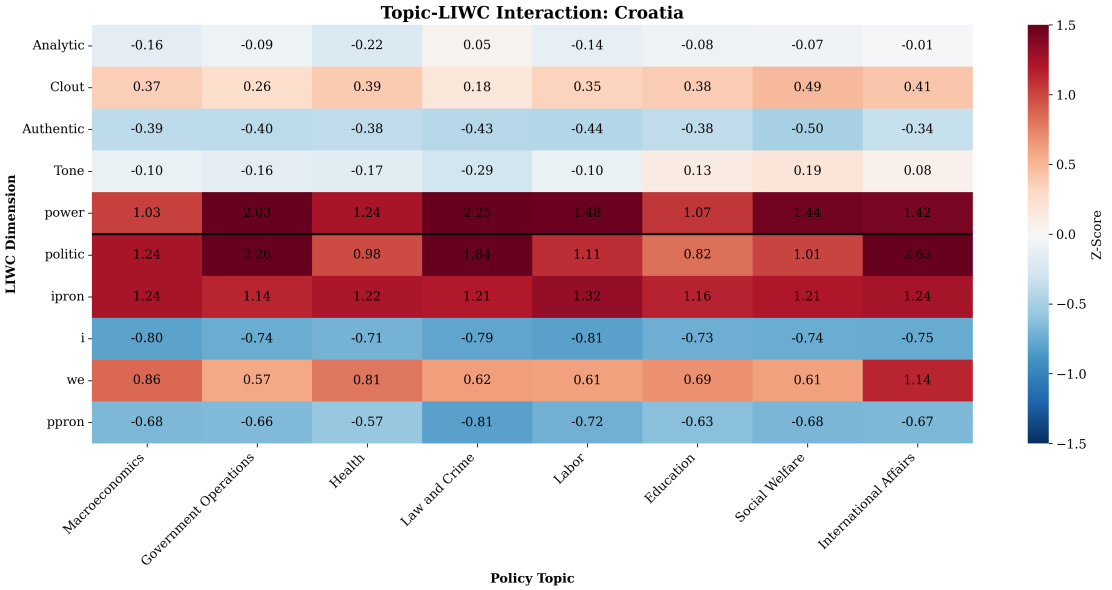


Figure 14.: **Topic–LIWC-22 interaction (Croatia).** Positive `politic` and `power` scores appear across domains, while the within-grid differences highlight topic-specific rhetorical emphases beyond this general parliamentary baseline.

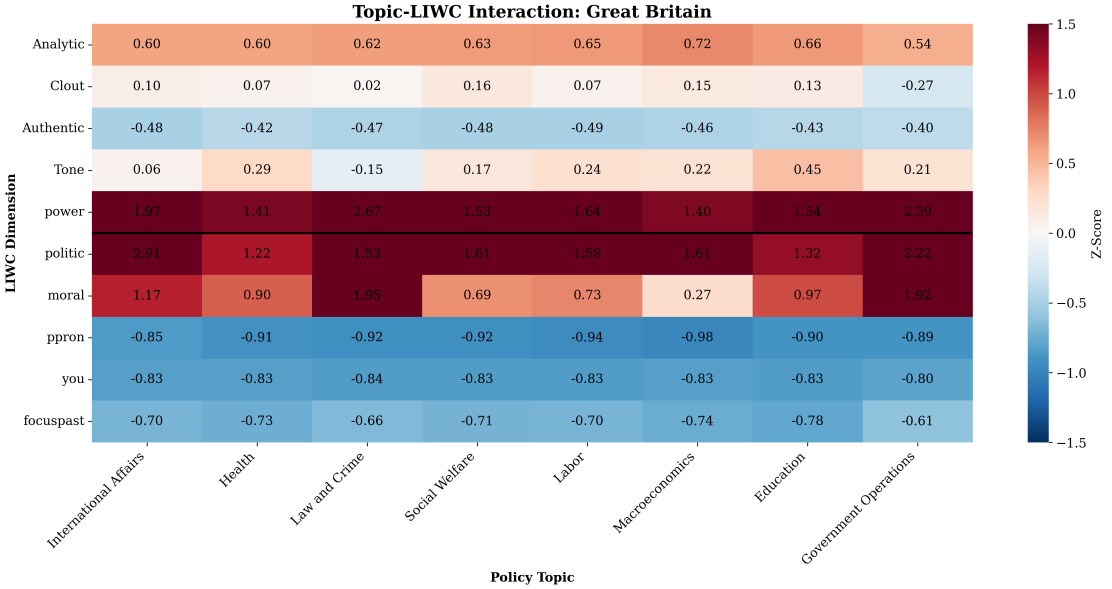


Figure 15.: **Topic–LIWC-22 interaction (Great Britain).** The grid shows strong political and power language across domains, with especially pronounced levels in international and security-related topics, and more social/narrative markers in welfare- and citizen-facing domains.

4.3. Ideology, Gender, and Age

4.3.1. Ideology

Across the left–right spectrum we see a broad "parliamentary" signature that is fairly stable: debates on all sides lean into power and political vocabulary, with a mild intensification toward the extremes. Figure 16 summarises this pattern. Each column is an ideological family, and rows are LIWC-22 dimensions.

The cells show group-mean LIWC z-scores relative to the Test Kitchen Corpus, not differences between ideological groups. This is why `politic` and `power` are strongly positive for *all* families: all parliamentary speeches are much more political and power-oriented than the general texts in the LIWC reference corpus. Differences between columns therefore indicate how much each family departs from an already highly politicised parliamentary baseline, rather than from zero.

Overall, ideological differences are structured but modest compared to topic and role effects, echoing earlier findings that institutional incentives and agenda choices often dominate partisan language (Grimmer et al. 2013; Proksch et al. 2010). The left–right divide matters, but less than what is being discussed and whether a party is in government or opposition.

4.3.2. Gender

Gender contrasts are small. Women tend to show slightly more positive tone, more work- and certainty-related language, and marginally lower emphasis on political and money vocabularies. Power language is essentially balanced. Country-specific idiosyncrasies exist—for example, Croatian women show somewhat higher certainty and work talk than Austrian women—but nothing overturns the general picture of subtle rather than dramatic gender gaps in parliamentary style.

Figure 17 summarises these patterns as female-minus-male z-score differences. Most cells lie in the interval $[-0.3, 0.3]$, and signs fluctuate across countries. There is no single, dominant "female style"; instead we see small, context-dependent shifts on top of the shared parliamentary baseline. This finding cautions against strong claims about gendered language in parliaments based solely on dictionary scores.

4.3.3. Age

Age patterns are clearer. Older MPs favour parliamentary and political language, speak more in social terms, and rely less on first-person singular and authenticity cues. Younger

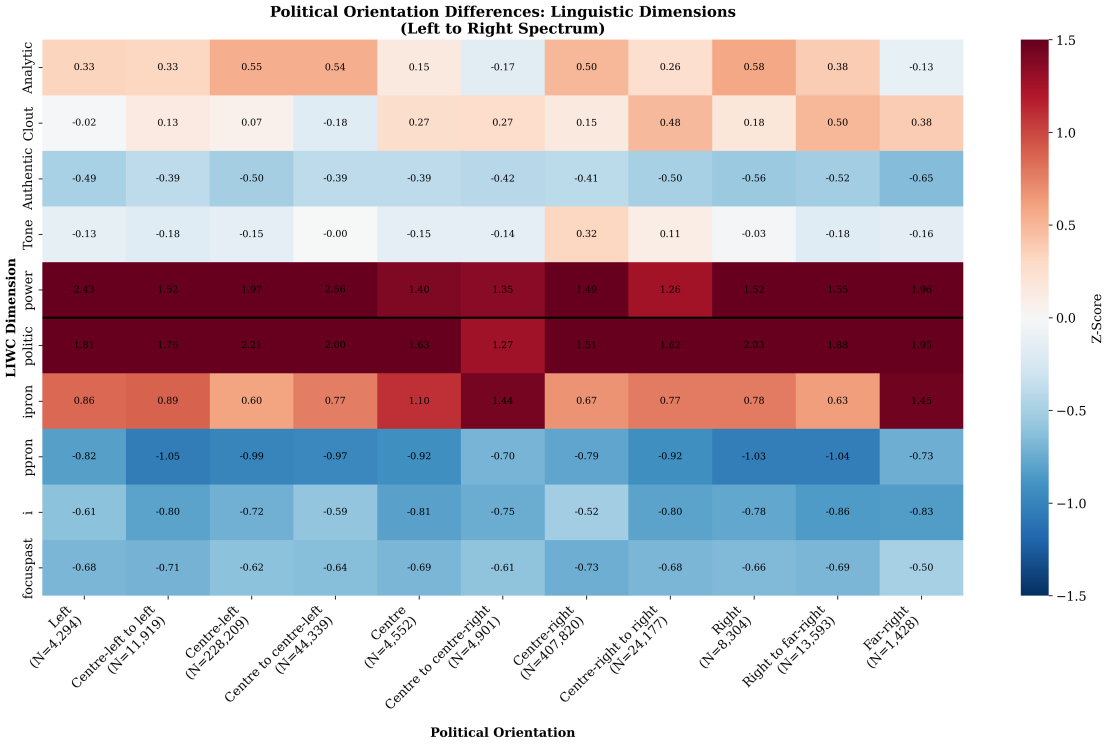


Figure 16.: **Ideological orientation.** Group-mean LIWC z-scores by ideological family, standardised relative to the LIWC-22 Test Kitchen Corpus. Each column is an ideological family, and rows are LIWC-22 dimensions. Left and right blocs share a common "parliamentary" profile (high `politic`, low `i`) but differ modestly on analytic and authenticity markers. These differences are smaller than topic and role effects, which supports the interpretation that institutional position and policy domain matter more for style than simple left–right placement.

cohorts display more certainty language and a somewhat more financial/transactional flavour. Analytic style edges upward with age, while overall tone remains steady.

Figure 18 groups speakers into four age bands and plots group-mean z-scores minus the overall mean. The oldest cohort (66+) is especially high on `power`, `politic`, and pronouns other than `i`, suggesting a strongly institutional and outward-facing register. By contrast, younger MPs show relatively higher `i` and *Authentic*, as well as somewhat stronger `money` vocabulary. These gradients suggest a life-cycle pattern in which parliamentary language becomes more institutional and less self-focused as careers progress. Because all values are still expressed as LIWC z-scores, levels well above zero (for example in `politic` and `power`) again reflect that all age groups use more of this vocabulary than the general-population texts in the Test Kitchen Corpus.

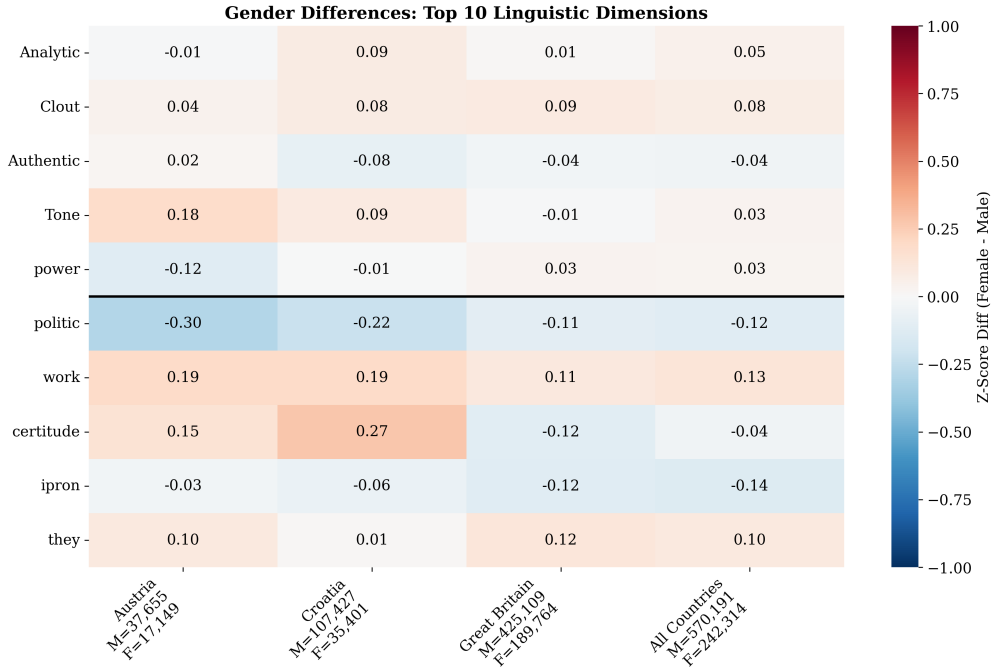


Figure 17.: **Gender contrasts** (female minus male). Warm (red) cells denote LIWC-22 categories that women use more than men, cool (blue) cells those that men use more. The overall pattern is one of small and heterogeneous effects: women display slightly higher positive tone and work/certainty talk and slightly lower money/politic scores in several panels, but there is no single, dominant "female style." These differences are weaker and less consistent than topic and role effects.

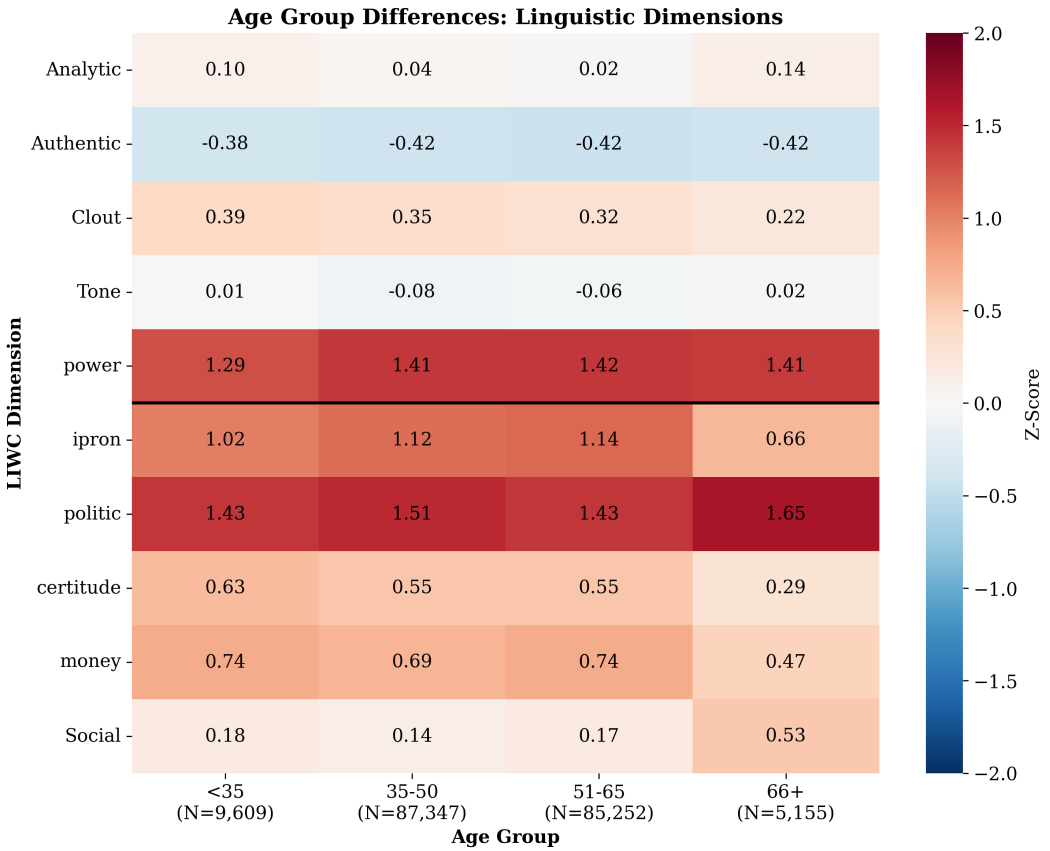


Figure 18.: **Age-group differences.** Group-mean LIWC z-scores by age band (centred on the overall parliamentary mean), still expressed on the LIWC-22 Test Kitchen z-scale. Columns separate age cohorts and rows show LIWC-22 dimensions. Older MPs talk more in institutional, social, and political terms, whereas younger MPs rely somewhat more on certainty language and money-related vocabulary. First-person singular and authenticity markers are relatively higher among younger speakers. These gradients suggest a life-cycle pattern in which parliamentary language becomes more institutional and less self-focused as careers progress.

4.4. Temporal Dynamics of Agenda and Style

Agendas and language respond to external events and shocks. We track LIWC-22 markers and topic prevalence with three-month moving averages. Shaded regions mark parliamentary terms. Vertical lines mark key external events.

Plot conventions. Alternating background shading shows successive parliamentary terms. A change of shading marks an election. Vertical dashed lines indicate major events.

4.4.1. Party Status Tone

Party-level illustration (Austria). Figure 19 reports four major Austrian parties (SPÖ, FPÖ, ÖVP, Grüne) and confirms a clear coupling between *institutional position* and tone. Across parties, governing periods are associated with warmer language, while opposition periods are cooler and more critical.

For the SPÖ, tone lifts sharply on entering government and steps down on moving to opposition, with the late-series opposition stretch showing a sustained flattening and intermittent dips below the overall mean. Short shocks (e.g., crisis episodes) appear as brief pullbacks even during government, but the level remains above opposition baselines.

For the FPÖ, the same status–tone linkage is visible but more volatile: coalition entries bring fast uplifts that decay sooner than for the SPÖ, and reversals on exiting government are abrupt.

For the ÖVP, status is essentially constant (coalition throughout the sample; panel header: 293 months coalition, 0 months opposition). Without role switches, we do not observe sharp step changes; instead, tone varies within a relatively narrow positive band, with only modest cyclical drift around crises and leadership changes.

For the Grüne, the series is dominated by opposition (238 months) with a short governing window at the end (31 months). The transition into government coincides with a visible upward level shift from a neutral/negative baseline to a clearly warmer register—consistent with the status effect observed for SPÖ and FPÖ.

Takeaway. Adding ÖVP and Grüne reinforces that tone tracks *party status* more than calendar time. Where status flips (SPÖ, FPÖ, Grüne), the three-month smoothing reveals discrete level shifts rather than transient blips. Where status is constant (ÖVP), tone shows bounded variation without regime breaks.

Other party-level series for *Analytic*, *Authentic*, *moral*, and emotion subdimensions appear in the appendix.

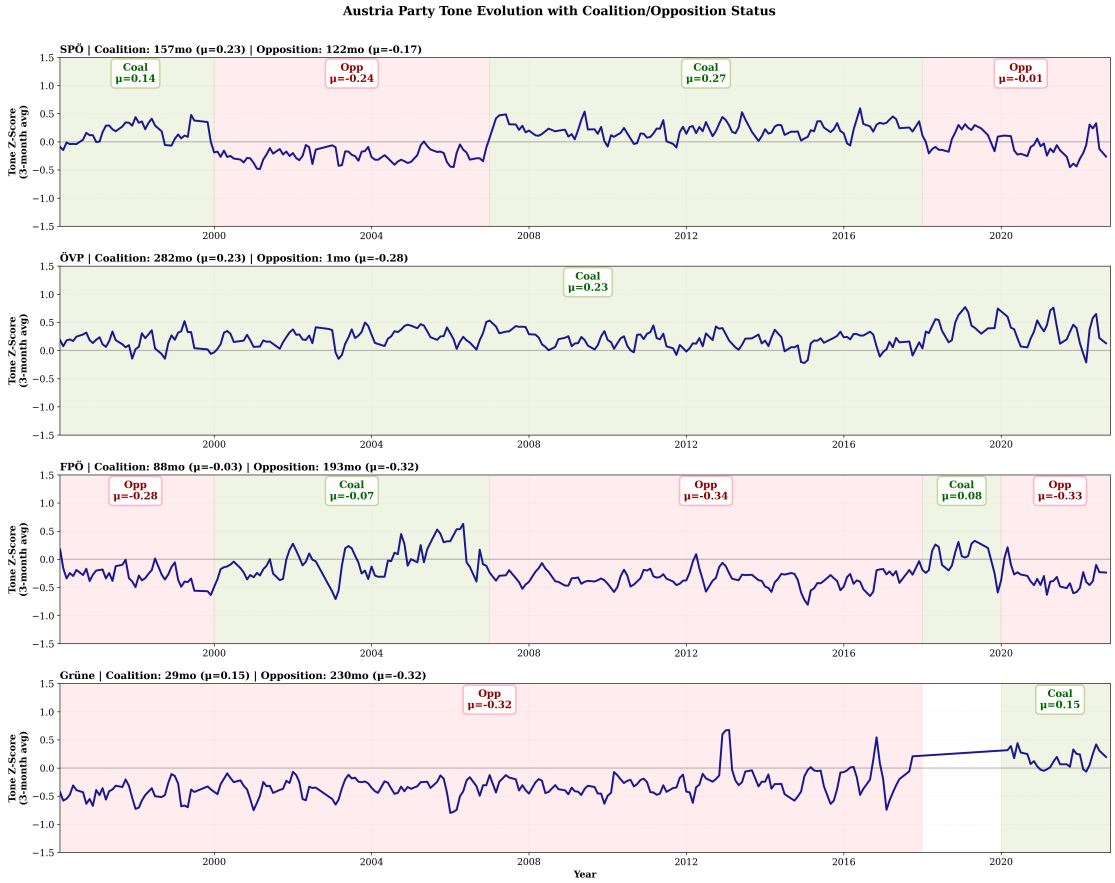


Figure 19.: **Austria party tone over time (SPÖ, FPÖ, ÖVP, Grüne).** Three-month moving averages of LIWC-22 Tone by party, with background shading indicating each party's own periods in government (green) and opposition (red). Panel headers report total months in each status. The figure shows clear level shifts when parties move between government and opposition, especially for SPÖ, FPÖ, and Grüne. In contrast, ÖVP spends the entire period in government and exhibits no comparable regime breaks.

4.4.2. Political Rhetoric Over Time

Figure 20 shows a common pattern across parliaments: the `politic` marker reacts sharply to shocks and then mean-reverts, while pronouns evolve more slowly. The collective register (`we`) trends upward over the long run, whereas `i` remains uniformly low. Direct address (`you`) oscillates with institutional routines (e.g., interpellations, Question Time) and is temporarily muted during periods of procedural disruption.

Austria. Political rhetoric steps up at major events and fades back toward baseline, with `we` higher and steadier during governing phases. `you` stays comparatively subdued,

consistent with fewer extended interrogation blocks.

Croatia. Series show clear regime shifts: `politic` spikes around the migration crisis and then normalises; `we` and `you` both lift around institutional changes associated with EU entry, suggesting more interactional formats in the chamber. Pandemic conditions temporarily dampen overt political language.

Great Britain. An early high-politics phase tied to the constitutional agenda gives way to a lower, flatter profile after the pandemic, with `we` broadly stable to rising and `you` suppressed during remote/modified sittings. The first-person singular remains minimal.
Takeaway. Event shocks produce level shifts in political rhetoric; institutional calendars modulate address terms. Over time, parliaments move toward a more collective voice (`we`) without a corresponding rise in self-reference, while acute crises temporarily amplify political vocabulary before it reverts.

4.4.3. Cognitive and Temporal Focus Over Time

Beyond political vocabulary and pronouns, we also track how MPs talk about thinking and time.

Cognitive processes. Figure 21 shows that cognitive-process markers (`cogproc`, `insight`, `cause`, `certain`) move more gradually than overt political rhetoric. Insight and causal language tend to rise around major policy reforms and during early pandemic months, when uncertainty is high and governments justify new measures. Certainty terms spike in moments of strong messaging—for example, in early crisis responses and during budget presentations—before returning toward baseline.

Temporal focus. Figure 22 shows that the present tense dominates parliamentary speech, but past- and future-focused language move in interpretable ways. Past focus increases around inquiries and retrospectives (e.g., post-crisis evaluations), while future focus rises in the run-up to elections and during strategic agenda-setting phases. These patterns again support the idea that our segment-level topics capture meaningful shifts in the temporal orientation of debates, not just noise in verb morphology.

Figure 20.: **Political language markers over time** (`politic`, `power`, `moral`, `money`). Each panel plots three-month moving averages for one country. Vertical lines mark major external shocks (e.g., migration crisis, COVID-19, Ukraine war); alternating background bands denote parliamentary terms. `politic` spikes during crises and constitutional moments and slowly mean-reverts, while `money` and `power` follow more cyclical patterns tied to budget and institutional debates. `moral` is comparatively stable. The figure summarises our core temporal finding for political rhetoric: shocks raise the temperature of debate, but the chamber does not remain permanently in "crisis mode."

Figure 21.: **Cognitive-process markers over time** (*Cognitive Processes*, *Insight*, *Cause*, *Certitude*). Three-month moving averages for each country. Cognitive-language scores respond less sharply than overt political markers: they drift upward during periods of intense policy justification (e.g., early pandemic months) and then stabilise. This suggests that parliamentarians adjust how much they explain, justify, and express certainty as the policy environment changes.

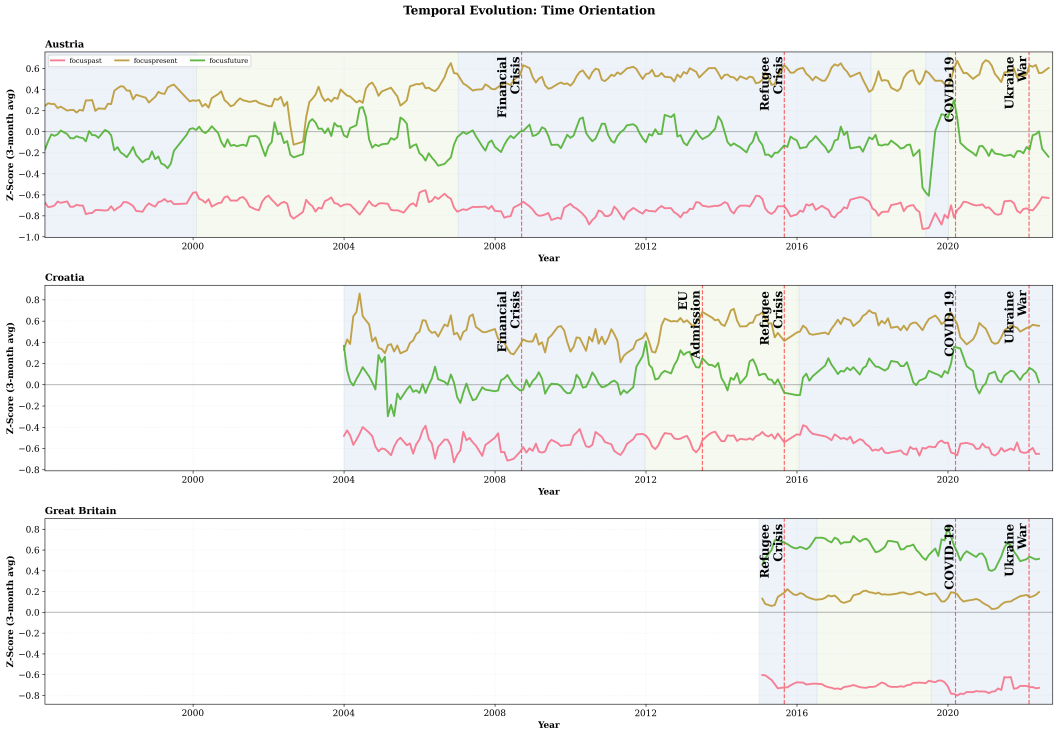


Figure 22.: **Temporal focus over time** (*Focus Past*, *Focus Present*, *Focus Future*). Present-focused language dominates throughout, but past- and future-oriented speech move systematically: past focus rises around inquiries and post hoc evaluations, while future focus increases in pre-election and strategy-heavy periods. The figure shows that our pipeline can recover not only which topics are on the agenda but also how far debates look backwards or forwards in time.

4.4.4. Economic vs. Health Topics

Figure 23 shows a clear substitution around the pandemic: as health policy dominates the agenda, macroeconomic debate recedes and only gradually recovers. Outside crisis periods, Macroeconomics exhibits regular, short-lived peaks that align with budget cycles—most visible in AT and GB—while Health tends to be flatter with longer plateaus once it rises. Croatia follows the same substitution logic but with a more sustained post-2020 health presence. Overall, shocks reallocate attention between these two domains rather than simply adding volume: when Health goes up, Macroeconomics typically goes down, and vice versa.

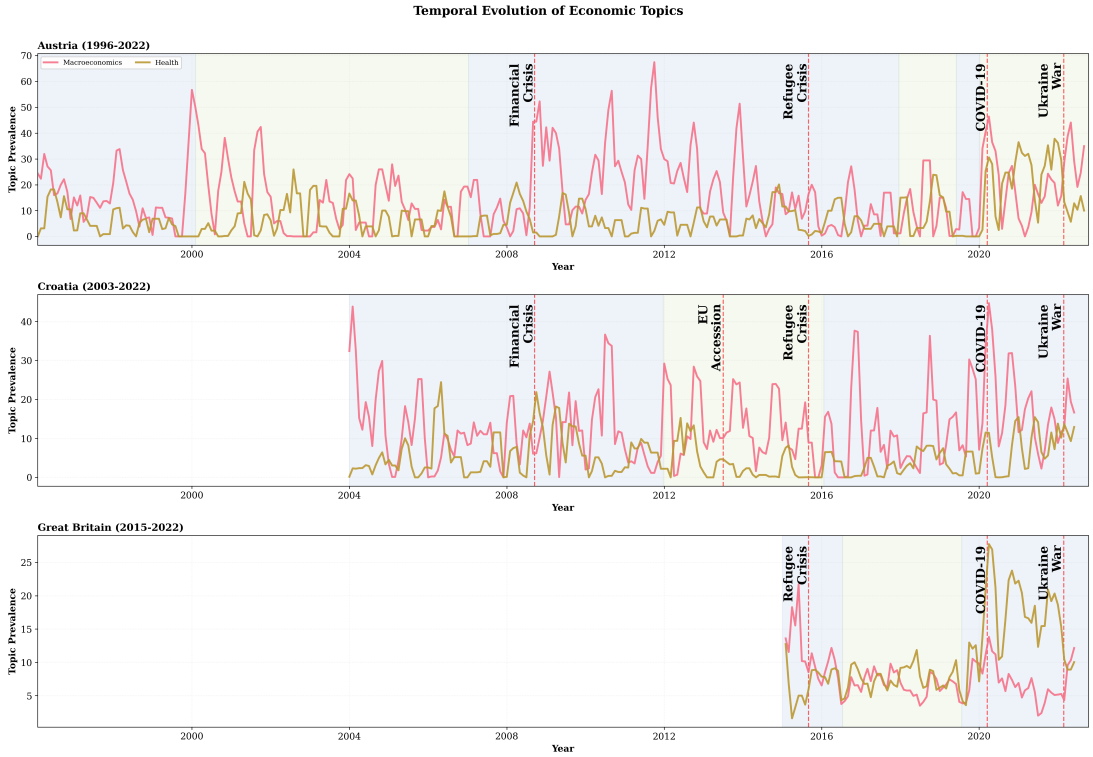


Figure 23.: **Economic vs. health topics over time.** Monthly prevalence of *Macroeconomics* and *Health* segments in AT, HR, and GB, smoothed with a three-month moving average. Alternating background bands mark parliamentary terms, and vertical lines indicate major crises such as COVID-19. In all three countries, the onset of the pandemic produces a clear substitution pattern: Health surges while Macroeconomics temporarily recedes, before both series drift back toward pre-crisis baselines. Budget cycles are visible as recurring bumps in Macroeconomics, especially in AT and GB.

4.4.5. Security and International Affairs

International Affairs tracks exogenous shocks: migration-related spikes in 2015–16 and a pronounced surge around the Ukraine war are visible across countries (Figure 24). Defence is comparatively stable and moves in smaller steps—procurement and force-posture debates appear as shorter bumps rather than long waves. Great Britain shows the sharpest International Affairs lift in the late period; Austria and Croatia register multiple smaller episodes rather than a single long cycle.

Figure 24.: **Security topics over time.** Prevalence of *International Affairs* and *Defence* segments across AT, HR, and GB (three-month moving averages). Shaded bands mark parliamentary terms; dashed vertical lines highlight key external events such as the 2015–2016 migration crisis and the Russian invasion of Ukraine. International Affairs rises sharply around these shocks and then partially recedes, whereas Defence shows smaller, more episodic increases linked to procurement decisions or force deployments.

Additional panels. Appendix C adds social and sustainability-related topics, which follow similar crisis- and election-linked dynamics.

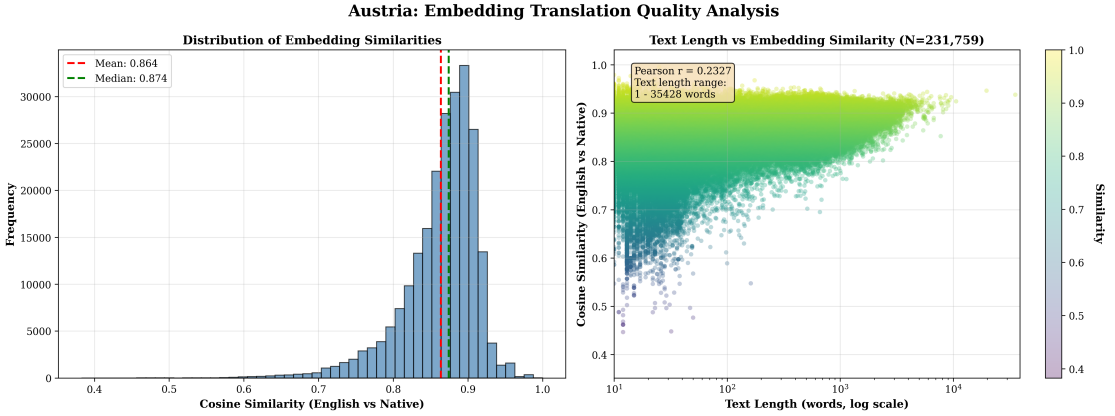
4.5. Cross-Lingual Embedding Consistency

For Austria and Croatia, native and MT-English embeddings of the same speech are very close in embedding space. For each utterance we compute a `BAAI/bge-m3` vector for the native text and another for the ParlaMint English translation, then take cosine similarity between the two. This produces one similarity score per speech and allows us to treat MT quality and multilingual alignment in a unified way.

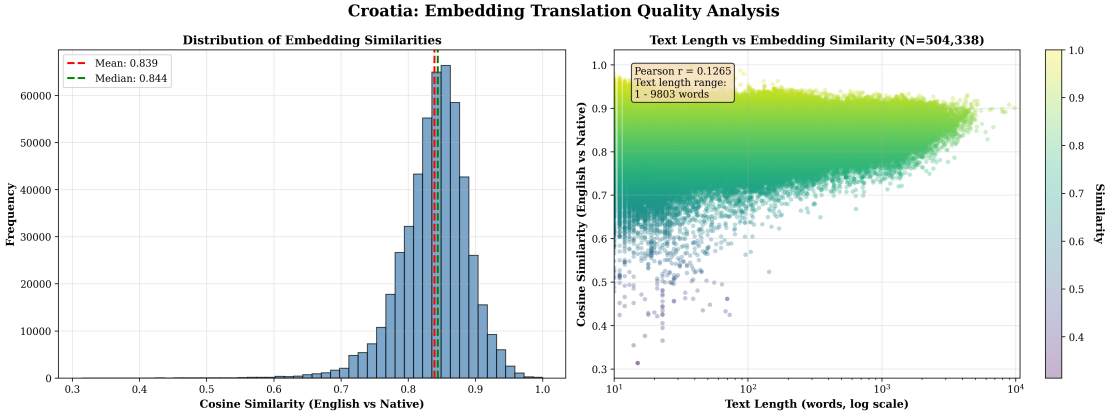
The left panels in Figure 25 show the empirical distributions of these similarities for AT and HR. Both histograms are sharply concentrated around high values: the mean is 0.864 for Austria and 0.839 for Croatia, with medians slightly above the means. The shape is close to Gaussian with light tails, and there is no sizeable mass of very low-similarity pairs. This suggests that only a small minority of texts suffer from poor alignment between native and translated versions. Put differently, for most speeches the multilingual encoder treats the English translation as a near-duplicate of the native text.

The right panels plot similarity against text length (log tokens). In both countries, similarity increases with length (Austria: Pearson $r = 0.233$; Croatia: $r = 0.127$, both $p < 0.001$). This is expected: longer speeches provide more lexical evidence for the encoder to lock onto, reducing the impact of local translation divergences or idioms. Very short contributions (e.g. interjections, procedural remarks) show more scatter, but these constitute a minor share of tokens and are down-weighted in segment-level aggregation. Importantly, there is no visible threshold at which long speeches would suddenly become unreliable—if anything, similarities for very long texts cluster at the upper end of the range.

Taken together, these diagnostics justify using a single English-based LIWC-22 pipeline for AT and HR with country-wise normalisation. The embedding model effectively treats native and translated texts as two noisy renderings of the same underlying semantic content, and length-dependent variation is modest relative to the overall level of alignment.



(a) Austria (AT): Mean=0.864, r=0.233



(b) Croatia (HR): Mean=0.839, r=0.127

Figure 25.: **Cross-lingual embedding consistency.** For each speech we embed both the native text and the ParlaMint English translation and compute cosine similarity between the two vectors. Left panels show the distribution of these similarities: most speeches cluster well above 0.8, indicating that the multilingual encoder treats native and translated texts as near-duplicates in embedding space. Right panels show that similarity increases mildly with length (log tokens), which is expected as longer texts provide more lexical evidence. Together, these diagnostics justify using a single English-based LIWC-22 pipeline for AT and HR with country-wise normalisation.

5. Discussion

5.1. Substantive Mechanisms Behind Stable Patterns

Our results show that *what* is discussed (policy domain) and *who* speaks (institutional role and demographics) are the dominant drivers of parliamentary style. Macroeconomics exhibits high `money` scores and slightly elevated *Clout*, paired with lower `moral`, `insight`, and `power` scores relative to other domains. Health debates show lower `politic`, `power`, and `money`, alongside mildly higher *Tone*. Coalition speech is associated with higher *Tone*, higher `we`, lower `you`, and lower `politic/power`; opposition speech shows the mirror pattern. These regularities are descriptive but stable across three parliaments.

This pattern fits classic accounts of political communication and parliamentary incentives. Governments face a joint task of policy-making and legitimacy maintenance; opposition parties face pressure to differentiate and scrutinise (Grimmer et al. 2013; Proksch et al. 2010). The coalition signature—*Tone*↑, `we`↑, `politic/power`↓—is consistent with the need to project competence and unity, while the opposition signature matches blame-attribution and conflict-oriented messaging (Lauderdale et al. 2016; Rheault et al. 2016). Topic-level differences also echo earlier work: economic debates tend to be number-heavy and focused on financial details (Slapin et al. 2008), whereas welfare and health debates contain more social and affective language (Del Gennaro et al. 2020; Rudkowsky et al. 2018).

Importantly, these effects persist after conditioning on topic. Within Macroeconomics or Health, coalitions still speak more positively and collectively than oppositions. This supports an interpretation in which role-based incentives shape style *within* domains, rather than merely reflecting different topic portfolios across parties.

5.2. Country-Specific Nuance Without Loss of Comparability

Cross-national regularities dominate our results, but we do observe interpretable deviations. GB economic debates show a mild `politic`↑ relative to AT/HR, consistent with Westminster's adversarial format and budget rituals that foreground partisan ownership.

Temporal profiles likewise differ in degree, not direction: migration, COVID-19, and the Ukraine war trigger rises in International Affairs and `politic` markers everywhere, followed by partial mean reversion, but the magnitudes and durations differ.

Two design choices make these comparisons robust. First, we rely on LIWC-22 z-scores based on the Test Kitchen Corpus, which provides a common general-population benchmark for all countries (Boyd et al. 2022b; Tausczik et al. 2010). Second, we align topics to a shared CAP taxonomy (Baumgartner et al. 2019), which ensures that "Macroeconomics" or "Health" refer to comparable issue bundles across countries even when institutional details differ (Erjavec et al. 2024). The combination of a common policy codebook and shared LIWC benchmarks yields patterns that are both interpretable and methodologically defensible.

5.3. Why Sequence Matters for Interpretation

Treating debates as sequences rather than bags of utterances is not just a technical choice; it changes what we can plausibly infer from the data. Agenda items unfold over dozens of turns, and many interventions use generic parliamentary vocabulary (procedural phrases, meta-comments, acknowledgements) that only make sense relative to neighbouring speeches. Purely local classifiers, whether dictionary-based or neural, cannot see this context. They therefore tend to fragment long episodes into short stretches of alternating labels, especially near agenda boundaries and on generic turns.

Our segmentation-first approach addresses this by inheriting topic labels from surrounding context. If a speech is locally ambiguous but sits in the middle of an otherwise coherent episode, it receives the dominant segment label. This reduces *boundary noise*: instead of dozens of spurious topic switches inside a single budget debate, we obtain one long Macroeconomics segment punctuated by a few genuine transitions to other domains. For descriptive tasks such as agenda tracking, this behaviour is closer to how MPs and observers talk about "what parliament was discussing" than per-utterance labels.

The downside is a deliberate mismatch with standard evaluation setups. Human coders in ParlaCAP label speeches in isolation, and balanced test sets include many boundary-adjacent utterances. In that context, a context-aware system that inherits labels from adjacent turns will often disagree with local judgments, even when it preserves the episode correctly. We view these disagreements as *expected side effects* of a segmentation-first design rather than as evidence against its usefulness. Our evaluation section, including the running example in Table 2, makes this trade-off clear: we sacrifice per-utterance F1 to gain episode-level coherence, which is more appropriate for time-series, role comparisons,

and narrative analyses in political communication (Grimmer et al. 2013).

5.4. Reading Effect Sizes in LIWC-22

We report z-scores throughout, always relative to the LIWC-22 Test Kitchen Corpus. Interpreting these magnitudes requires some care. Because LIWC-22 categories count the proportion of tokens belonging to curated dictionaries, even small standardised shifts can be meaningful at scale (Pennebaker et al. 2015; Tausczik et al. 2010). In our context:

- Effects around $|0.1|$–$|0.3|$ mark subtle but systematic preferences in lexical choice (e.g., slightly more certainty language for younger MPs).
- Effects around $|0.3|$–$|0.6|$ indicate moderate, clearly interpretable differences in rhetorical emphasis (e.g., coalition vs. opposition gaps in `politic`).
- Effects above $|1.0|$ (such as the `money` score of Macroeconomics) correspond to strong, domain-defining patterns.

Summary variables (*Analytic*, *Clout*, *Authentic*, *Tone*) move less than content dictionaries, which is expected: they aggregate many micro-choices and are designed to be conservative and stable across contexts (Boyd et al. 2022b). Content dictionaries, by contrast, are closer to topic vocabularies and therefore show sharper contrasts between domains. This explains why Macroeconomics stands out so clearly on `money` while *Analytic* remains near zero, and why Health differs mainly in `politic`, `power`, and `money` rather than in a wholesale shift of all summary dimensions.

Because our sample is very large, almost any deviation from zero would be "statistically significant" in a hypothesis test. A difference of 0.3 in `moral`, for instance, means that one group uses moral dictionary words 0.3 standard deviations more often than the Test Kitchen benchmark. This is numerically small but highly consistent across millions of words. For this reason we emphasise effect sizes and patterns across countries rather than significance tests.

5.5. Robustness and Design Trade-Offs

Three modelling choices are especially consequential for our results: segmentation signals, dimensionality reduction and clustering, and CAP mapping.

Segmentation signals. Our hybrid segmentation combines semantic shifts with chair cues. Removing chair cues reduces precision at boundaries, particularly in settings like GB where explicit agenda phrases are rare. Conversely, increasing the weight on chair cues would under-detect semantic shifts that are not flagged procedurally. Our automatic window-size optimisation mitigates some of this tension by selecting window sizes that jointly maximise within-segment coherence and alignment with chair turns, but it does not resolve structural differences in how parliaments are chaired. This likely contributes to the slightly lower performance of GB on human tests and should be reconsidered when extending the approach to chambers with very different procedural conventions.

Dimensionality reduction and clustering. We reduce 1024-dimensional embeddings to 10 dimensions with UMAP before clustering with a GMM. Empirically, very low target dimensions merge nearby policy areas (e.g., conflating Health and Social Welfare), whereas very high dimensions reintroduce unstable, high-dimensional distance behaviour and make GMM fitting brittle (McInnes et al. 2018). Alternative pipelines—e.g., principal components followed by HDBSCAN or k-means—either produced too many outliers or unrealistically balanced clusters. Our choice of UMAP+GMM reflects a pragmatic trade-off between stability, interpretability, and computational tractability.

CAP mapping via LLM. Mapping unsupervised clusters to CAP domains is inherently uncertain. We use conservative prompts, deterministic decoding, and a one-label-per-cluster rule. This avoids overfitting to idiosyncratic keyword lists but hides genuine mixtures (e.g., combined economic and welfare reforms) and forces discrete decisions where human coders might assign multiple codes (Baumgartner et al. 2019). The decision to fall back to *Other* or *Mix* when uncertain prevents exotic mislabels but slightly inflates generic categories, especially for small or noisy clusters.

Overall, our robustness checks suggest that substantive patterns—Macroeconomics vs. Health, coalition vs. opposition, status-driven tone shifts, and crisis substitution—remain stable under plausible parameter changes. Where results do change, they mostly affect the fine structure of less frequent domains rather than the core regularities highlighted in the main figures.

5.6. Limitations and Implications for Evaluation

Our approach has several limitations that should be kept in mind when interpreting the findings or re-using the pipeline.

Context vs. local accuracy. By design, we optimise for episode-level coherence rather than per-utterance accuracy. This means that standard F1 metrics on isolated speeches will understate performance for the tasks we actually care about (segment-level agenda tracking, topic prevalence over time). Human coders also show only moderate agreement on CAP labels (*ParlaCAP: Comparing agenda settings across parliaments via the ParlaMint dataset* 2025), and disagreements are especially common near boundaries or for generic interventions. Future evaluations should therefore complement utterance-level metrics with segment-level boundary accuracy, run-length coherence, and measures that explicitly reward stable episode labelling.

Machine translation and cross-lingual comparability. For AT and HR we rely on ParlaMint English translations to apply LIWC-22, because English is the best-supported language in the instrument (Boyd et al. 2022a). Our cross-lingual embedding checks indicate high alignment (mean cosine >0.83) between native and translated texts, but subtle nuances—especially wordplay, idioms, or culturally specific references—may still be lost in translation. This could dampen or distort some stylistic signals, particularly in affective and moral categories (McDonnell et al. 2020). We partially mitigate this through within-country normalisation of segment averages and by focusing on relatively robust, high-frequency categories, but the limitation remains.

Dictionary and measurement biases. LIWC-22 is a transparent, interpretable instrument, but its dictionaries were developed primarily on English-language and often US-centric data (Boyd et al. 2022b; Tausczik et al. 2010). Even after translation, certain concepts may map differently onto political discourse in Austria, Croatia, and Great Britain than in the contexts where the dictionaries were validated. For example, the `moral` or `power` categories may capture different argumentative traditions in Catholic vs. Protestant or post-socialist settings. Our results should therefore be read as stylised indicators of relative emphasis, not as definitive psychological diagnoses of individual MPs or parties.

Coverage and institutional focus. We analyse three European parliaments over limited time windows. While they provide variation in language, party systems, and institutional design, they are not globally representative. For example, we do not cover presidential or strongly majoritarian systems outside Europe, upper chambers, or committee-level deliberations, all of which may exhibit different patterns of topic structure and rhetorical style (Lauderdale et al. 2016). Extending the analysis to more countries and institutional

settings is an important next step.

Downstream use and ecological validity. Our segment labels and stylistic profiles are best suited for aggregate analyses: time-series of topic prevalence, comparative dashboards, and role-based contrasts. Using them to make claims about individual MPs, short time windows, or causal effects on policy outcomes requires additional assumptions. In particular, we do not estimate the *impact* of language on vote outcomes or public opinion; we only describe how language systematically varies by topic, role, and time. Any downstream application that moves from description to prediction or intervention should therefore re-validate the measures in the specific task environment (Grimmer et al. 2013).

5.7. Practical Implications for Stakeholders

Despite these limitations, the pipeline provides concrete value for different stakeholders.

Parliamentary administrations and research services. For internal users, the segmentation and topic-labelling framework can underpin dashboards that track agenda composition in near real time. Instead of monitoring thousands of speeches, staff can follow segment-level CAP series and surface episodes where, say, Health or International Affairs suddenly spike. Combined with LIWC-22 profiles, this enables quick diagnostics of whether crises are associated with more confrontational language, more collective framing, or shifts in temporal focus. Such tools can support strategic planning, communication, and evaluation of parliamentary reforms.

Parties and speechwriters. Political actors can use the domain- and role-specific style profiles to calibrate their messaging. For example, when a party enters government and wants to signal responsibility and competence, our findings suggest that a more collective, positive, and less overtly political register aligns with historical incumbency patterns. In budget debates, emphasising clear, concrete money vocabulary is consistent with the Macroeconomics profile we observe, whereas injecting moral rhetoric would represent a deliberate departure. Conversely, oppositions seeking to sharpen their scrutiny can lean into the more adversarial patterns (`politic`, `power`, `you`) characteristic of challengers, while remaining aware of potential polarisation risks.

Business stakeholders (investors, firms, and risk functions). For asset managers, banks, consultancies, and corporate risk teams, the pipeline can act as a structured

layer for **policy-risk intelligence**. Segment-level CAP time series can be used to monitor when attention to regulation-relevant domains (e.g., Energy, Environment, Labor, Macroeconomics, Health, Foreign Trade) rises sharply or shifts across parties and roles, enabling earlier identification of emerging regulatory pressure or fiscal priorities. Combining topic prevalence with LIWC-22 style markers further supports triage: spikes in political vocabulary, conflict markers, or certainty language can flag moments when negotiations intensify or when governments attempt to stabilise expectations. In practice, these outputs are most useful as *decision-support signals*—to prioritise analyst attention, support scenario planning, and document risk narratives—rather than as standalone predictors of market outcomes.

Journalists, NGOs, and civic-tech projects. External observers can re-use our segment-level labels and stylistic metrics to contextualise individual debates. Instead of cherry-picking colourful quotes, coverage can situate discourse within its broader episode: Was this an unusually conflictual debate for this domain? Did the government depart from its typical incumbent style? For NGOs and civic-tech initiatives, the pipeline can support public dashboards that track agenda attention to issues such as climate, migration, or welfare, combined with simple visualisations of tone and collective framing.

Comparative researchers. Scholars in comparative politics and political communication can treat our pipeline as an intermediate layer: a way of turning raw parliamentary corpora into a structured, cross-nationally comparable representation of topics, episodes, and styles. This can feed into studies of representation, responsiveness, coalition dynamics, or the link between parliamentary discourse and public opinion (Lauderdale et al. 2016; Proksch et al. 2010). Because the method is weakly supervised and relies on standard components, it can be adapted to other parliaments with relatively low engineering overhead.

5.8. Where This Leaves Supervised Modelling

Our segmentation-based labels are not a competitor to fully supervised systems trained on large, high-quality annotation sets. Instead, they are a source of *structure and pseudo-labels* that supervised models can exploit.

One natural extension is to treat segment labels as soft constraints in a sequence model. For example, utterance-level classifiers (e.g., transformers fine-tuned on ParlaCAP labels) could be combined with a conditional random field or HMM that penalises frequent

topic switches within segments and encourages transitions at detected boundaries (Blei et al. 2006; Gruber et al. 2007). Alternatively, segment labels could be used as weak supervision in self-training or distillation schemes, where a model is trained to approximate the segment-informed labels while having access to full utterance context.

A second avenue is to jointly learn segmentation and topic assignment, for example by using neural architectures that predict both boundary and topic variables and incorporate constraints from chair cues and metadata (Koshorek et al. 2018). Our work suggests that even simple, geometry-based segmentation already yields useful structure; learning segmentation directly from text and labels may further improve both boundary accuracy and topic coherence.

In both cases, the key lesson is that supervision should reflect the sequential nature of debates. Purely local labels, whether human- or model-generated, will always struggle with generic and boundary-adjacent speeches. Sequence-aware supervision—whether via segments, weak constraints, or joint models—offers a way to reconcile the strengths of supervised learning with the interpretability of episode-based analysis.

6. Conclusion

Summary of contributions. This thesis develops and applies a sequence-aware pipeline for analysing parliamentary debates across three European countries. Methodologically, we segment sittings into coherent episodes using hybrid semantic and procedural signals, cluster segment embeddings, and map them to CAP policy domains with conservative LLM label assignment. Parallel LIWC-22 scoring provides psycholinguistic profiles by topic, role, and time. Substantively, we show that policy domains and institutional roles, rather than individual or party characteristics alone, are the dominant drivers of parliamentary style.

Substantive findings. Across Austria, Croatia, and Great Britain, Macroeconomics debates exhibit strong over-use of money vocabulary and slightly higher authority/Clout, but lower moral and insight language than other domains, consistent with a style focused on financial details. Health debates feature lower political and power markers and slightly higher tone, pointing to a less adversarial, more service-oriented rhetoric. Coalition speakers use more collective and positive language and less overtly political and power-related wording; oppositions show the reverse, sharpening conflict and direct address. Temporal analyses reveal predictable crisis substitution (Health vs. Macroeconomics during COVID-19) and short-lived spikes in political rhetoric around shocks such as the migration crisis and the war in Ukraine. Age and gender differences exist but are modest compared to topic and role effects.

Why sequence matters. A central message of this work is that parliamentary debates are better understood as sequences of agenda episodes than as collections of isolated speeches. Segment-first labelling reduces boundary noise, produces more interpretable topic runs, and yields time-series that align with institutional calendars and external shocks. While this comes at a cost in per-utterance F1 on context-free tests, it better matches how MPs, journalists, and citizens describe "what parliament was talking about" on a given day.

Implications for stakeholders. For parliamentary administrations and civic-tech projects, the pipeline provides a scalable way to monitor agenda composition and rhetorical style in near real time. For parties and speechwriters, it offers empirical benchmarks on how incumbents and oppositions typically speak within different policy domains, which can inform strategic message design. For researchers, it demonstrates that unsupervised, sequence-aware methods can recover stable, interpretable regularities across multiple parliaments and languages, opening up new possibilities for comparative text-as-data work.

Future work. Several extensions follow naturally. First, scaling to additional ParlaMint parliaments and beyond Europe would test the generality of our findings and identify regime- or culture-specific exceptions. Second, integrating learned segmentation and segment-aware supervised models could improve utterance-level performance without sacrificing coherence. Third, linking linguistic measures to legislative outcomes, media coverage, or public opinion would move from description toward causal explanations of how language shapes political behaviour and perceptions. Finally, richer stylistic instruments (e.g., contextual moral foundations, narrative structure) could complement LIWC-22 and capture dimensions of rhetoric that dictionary methods miss (Del Gennaro et al. 2020; Rudkowsky et al. 2018).

Outlook. Parliamentary transcripts are one of the most transparent records of democratic politics. By aligning topic modelling and psycholinguistic profiling with the sequential structure of debates, this thesis offers a more faithful lens on how parliaments allocate attention, frame issues, and adapt their rhetoric to changing circumstances. The hope is that this combination of scalable methods and interpretable measures can contribute to both scholarly understanding and practical tools for monitoring and improving democratic deliberation.

Bibliography

Abercrombie, Gavin and Riza Batista-Navarro (2020). "Sentiment and position-taking analysis of parliamentary debates: a systematic literature review". In: *Journal of Computational Social Science* 3.1, pp. 245–270.

– (2022). "Policy-focused Stance Detection in Parliamentary Debate Speeches". In: *Northern European Journal of Language Technology* 8.

Akaike, Hirotugu (1974). "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Alexander, Marc and Mark Davies (2015). *The Hansard Corpus 1803–2005*. Corpus website.

Alvarez, R. Michael and Jacob Morrier (2025). *Measuring the Quality of Answers in Political Q&As with Large Language Models*. arXiv: 2404.08816 [cs.CL].

Artetxe, Mikel and Holger Schwenk (2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *Transactions of the Association for Computational Linguistics* 7, pp. 597–610.

Bäck, Hanna and Marc Debus (2019). "When do Women Speak? A Comparative Analysis of the Role of Gender in Legislative Debates". In: *Political Studies* 67.3, pp. 576–596.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016). "Measuring Economic Policy Uncertainty". In: *The Quarterly Journal of Economics* 131.4, pp. 1593–1636.

Baumgartner, Frank R., Christoffer Green-Pedersen, and Bryan D. Jones, eds. (2013). *Comparative Studies of Policy Agendas*. Cambridge University Press.

Baumgartner, Frank R., Bryan D. Jones, and John Wilkerson (2019). *Policy Agendas Codebook (U.S.)* Comparative Agendas Project.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of ICML*. ACM, pp. 113–120.

Boyd, Ryan L., Ashokkumar Ashwini, and James W. Pennebaker (2022a). *LIWC-22 Psychometrics Manual*.

Boyd, Ryan L. et al. (2022b). *The Development and Psychometric Properties of LIWC-22*. Technical manual.

Caldara, Dario and Matteo Iacoviello (2022). “Measuring Geopolitical Risk”. In: *American Economic Review* 112.4, pp. 1194–1225.

Calinski, Tadeusz and Jerzy Harabasz (1974). “A Dendrite Method for Cluster Analysis”. In: *Communications in Statistics* 3.1, pp. 1–27.

Campello, Ricardo J. G. B., Davoud Moulavi, and Jörg Sander (2013). “Density-Based Clustering Based on Hierarchical Density Estimates”. In: *Proceedings of PAKDD*, pp. 160–172.

Campello, Ricardo J. G. B. et al. (2015). “Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection”. In: *ACM Transactions on Knowledge Discovery from Data* 10.1, 5:1–5:51.

Campos, Ricardo et al. (2020). “YAKE! Keyword Extraction from Single Documents using Multiple Local Features”. In: *Information Sciences* 509, pp. 257–289.

Chen, Jun et al. (2024). “M3-Embedding: Multi-Linguality, Multi-Functionality, and Multi-Granularity”. In: *Findings of ACL*.

Choi, Freddy Y. Y. (2000). “Advances in Domain Independent Linear Text Segmentation”. In: *Proceedings of NAACL*, pp. 26–33.

Davies, David L. and Donald W. Bouldin (1979). “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2, pp. 224–227.

Decter-Frain, Ari and Jeremy A. Frimer (2016). “Impressive Words: Linguistic Predictors of Public Approval of the U.S. Congress”. In: *Frontiers in Psychology* 7, p. 240.

Del Gennaro, Camilla et al. (2020). “Moral Foundations in Political Discourse: A Cross-Country Perspective”. In: *Political Communication*.

Eisenstein, Jacob and Regina Barzilay (2008). “Bayesian Unsupervised Topic Segmentation”. In: *Proceedings of EMNLP*, pp. 334–343.

Erjavec, Tomaž et al. (2024). “ParlaMint II: advancing comparable parliamentary corpora for a global community”. In: *Language Resources and Evaluation*.

Feng, Fuli et al. (2020). “Language-agnostic BERT Sentence Embedding”. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 911–920.

Fox, Emily B. et al. (2008). “Sticky HDP-HMMs for Sequence Data”. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1369–1376.

Grimmer, Justin and Brandon M. Stewart (2013). “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. In: *Political Analysis* 21.3, pp. 267–297.

Grootendorst, Maarten (2022). “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. In: *arXiv*. eprint: 2203.05794.

Gruber, Amit, Michal Weiss, and Michal Rosen-Zvi (2007). “Hidden Topic Markov Models”. In: *Proceedings of AISTATS*, pp. 163–170.

Hanretty, Chris, Michael Marsh, and Ben Stott (2025). *The Future Is Old: Aging and Representation in the House of Commons*. Preprint.

Hassan, Tarek A. et al. (2019). “Firm-Level Political Risk: Measurement and Effects”. In: *The Quarterly Journal of Economics* 134.4, pp. 2135–2202.

Hearst, Marti A. (1997). “TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages”. In: *Computational Linguistics*.

Hiltunen, Turo, Tony McEnery, and Arne Ziegler (2020). “Investigating colloquialization in the British parliamentary record”. In: *Language Sciences*.

Hirst, Graeme et al. (2014). “Argumentation and Ideology: An Analysis of Mainstream and Alternative News Media”. In: *Proceedings of the Workshop on Argumentation Mining*. CEUR-WS.

Jensen, Jacob et al. (2012). “Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech”. In: *Brookings Papers on Economic Activity* 43.2 (Fall), pp. 1–81.

Jordan, Kayla N. et al. (2019). “Examining long-term trends in politics and culture through the language of political leaders and cultural institutions”. In: *Proceedings of the National Academy of Sciences*.

Körner, Robert et al. (2022). “How the Linguistic Styles of Donald Trump and Joe Biden Reflect Different Forms of Power”. In: *Journal of Language and Social Psychology*, pp. 1–28.

Koshorek, Omri et al. (2018). “Text Segmentation as a Supervised Learning Task”. In: *arXiv*. eprint: 1803.09337.

Kuzman, Taja, Nikola Ljubešić, and Daniela Širinić (2025). *ParlaCAP: Annotation with CAP Topics*. GitHub repository.

Lauderdale, Benjamin E. and Alexander Herzog (2016). “Measuring Political Positions from Legislative Speech”. In: *Political Analysis* 24.3, pp. 374–394.

Laver, Michael, Kenneth Benoit, and John Garry (2003). “Extracting Policy Positions from Political Texts Using Words as Data”. In: *American Political Science Review* 97.2, pp. 311–331.

McDonnell, Mary, Elia Psouni, and Sven Bölte (2020). “Validation of Linguistic Inquiry and Word Count (LIWC) in emotional expression”. In: *Frontiers in Psychology* 11, p. 1590.

McInnes, Leland, John Healy, and James Melville (2018). “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: *arXiv*. eprint: 1802.03426.

Mihalcea, Rada and Paul Tarau (2004). “TextRank: Bringing Order into Texts”. In: *Proceedings of EMNLP*, pp. 404–411.

Monroe, Burt L., Michael Colaresi, and Kevin M. Quinn (2008). “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict”. In: *Political Analysis* 16.4, pp. 372–403.

ParlaCAP: Comparing agenda settings across parliaments via the ParlaMint dataset (2025). OSCARS project page.

ParlaMint 5.0 now available (July 2025). CLARIN ERIC News.

ParlaMint: Comparable Parliamentary Corpora (n.d.). GitHub repository.

Pástor, Ľuboš and Pietro Veronesi (2013). “Political Uncertainty and Risk Premia”. In: *Journal of Financial Economics* 110.3, pp. 520–545.

Pennebaker, James W. et al. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX.

Proksch, Sven-Oliver and Jonathan B. Slapin (2010). “Position Taking in European Parliament Speeches”. In: *British Journal of Political Science* 40.3, pp. 587–611.

Rauh, Christian (2020). *ParlSpeech V2 and related datasets*. Data and resources page.

Rauh, Christian and Jan Schwalbach (2020). *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches*. Release note.

Reimers, Nils and Iryna Gurevych (2020). “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Reynolds, Douglas A. (2009). “Gaussian Mixture Models”. In: *Encyclopedia of Biometrics*. Ed. by Stan Z. Li and Anil K. Jain. Boston, MA: Springer US, pp. 659–663.

Rheault, Ludovic et al. (2016). “Measuring Emotion in Parliamentary Debates with Automated Text Analysis”. In: *PLOS ONE* 11.12, e0168843.

Riedl, Martin and Chris Biemann (2012). “TopicTiling: A Text Segmentation Algorithm Based on LDA”. In: *Proceedings of the ACL 2012 Student Research Workshop*, pp. 37–42.

Rousseeuw, Peter J. (1987). “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”. In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65.

Rudkowsky, Emanuel et al. (2018). “More than Bags of Words: Sentiment Analysis with Word Embeddings”. In: *Austrian Journal of Political Science*.

Schwalbach, Jan et al. (2024). *ParlLawSpeech: Linked corpora of bills, laws, and plenary speeches*. Dataset website.

Schwarz, Gideon (1978). “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2, pp. 461–464.

Sebők, Miklós and Zoltán Kacsuk (2021). “The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach”. In: *Political Analysis* 29.2, pp. 236–249.

Slapin, Jonathan B. and Sven-Oliver Proksch (2008). “A Scaling Model for Estimating Time-Series Party Positions from Texts”. In: *American Journal of Political Science* 52.3, pp. 705–722.

Sylwester, Karolina and Matthew Purver (2015). “Twitter Language Use Reflects Psychological Differences between Democrats and Republicans”. In: *PLOS ONE* 10.9, e0137422.

Tausczik, Yla R. and James W. Pennebaker (2010). “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods”. In: *Journal of Language and Social Psychology* 29.1, pp. 24–54.

Wang, Xuerui and Andrew McCallum (2006). “Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends”. In: *Proceedings of KDD*, pp. 424–433.

Zhang, Justine, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil (2017). “Asking too much? The rhetorical role of questions in political discourse”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1558–1572.

A. Human Agreement on CAP Labels

Table 6 reproduces nominal Krippendorff's α reported by the ParlaCAP team for pairwise agreement between three human annotators and GPT-4o (*ParlaCAP: Comparing agenda settings across parliaments via the ParlaMint dataset* 2025). These values illustrate that even expert coders and state-of-the-art models disagree on a substantial fraction of speeches, especially at topic boundaries, which motivates more modest expectations for unsupervised, segment-level models.

Table 6.: Pairwise nominal Krippendorff's α between human annotators and GPT-4o on the ParlaCAP human test sets.

Pair	**Krippendorff's α**
Ann1 & Ann3	0.678
GPT-4o & Ann3	0.642
GPT-4o & Ann1	0.633
Ann2 & Ann3	0.618
GPT-4o & Ann2	0.600
Ann1 & Ann2	0.591

B. LLM Classification Prompt

The following prompt was used to classify parliamentary speech topics into Comparative Agendas Project (CAP) categories. It was supplied as the user message, together with a fixed system message, to the GPT-4o-mini model.

Prompt text

```
System message:
"You are a parliamentary policy classifier. Always respond in English."

User message:

Analyze these parliamentary keywords and provide TWO outputs IN ENGLISH:

Country: [Country] Parliament
Source Language: [Language]
Keywords: [Top 15 n-grams from topic model]

TASK 1 - Topic Name:
Create a short, descriptive name IN ENGLISH (2-4 words) that captures what
this topic is about as discussed in the parliamentary meeting.

TASK 2 - CAP Classification:
Classify into ONE of these policy categories:

Education, Technology, Health, Environment, Housing, Labor, Defense,
Government Operations, Social Welfare, Macroeconomics, Domestic Commerce,
Civil Rights, International Affairs, Transportation, Immigration,
Law and Crime, Agriculture, Foreign Trade, Culture, Public Lands,
Energy, Other, Mix

Instructions:
- Always respond in English, even if keywords are in German, Croatian,
```

```
  or other languages.
- For Topic Name: be specific and descriptive (e.g., "Healthcare Reform",
  "Military Defense Budget").
- For CAP Classification: choose the most specific policy category.
- Use "Other" for procedural / non-policy content.
- Use "Mix" only if the topic clearly spans multiple domains.
- Be conservative: default to "Other" if uncertain.

Format your response EXACTLY as:
TOPIC: [your English topic name]
CATEGORY: [exact category name from list]
```

Model configuration

- Model: GPT-4o-mini
- Temperature: 0.00
- Max tokens: 300

C. Additional Temporal Analyses

Figure 26.: **LIWC-22 summary variables over time** (*Analytic*, *Clout*, *Authentic*, *Tone*). Three-month moving averages by country. The series show that parliamentary style is relatively stable in the long run, with temporary deviations around major shocks and government changes.

Figure 27.: **Affect over time** (overall affect, positive tone, negative tone). Positive and negative tone move asymmetrically around crises: negative tone rises sharply in acute phases, while positive tone recovers more slowly.

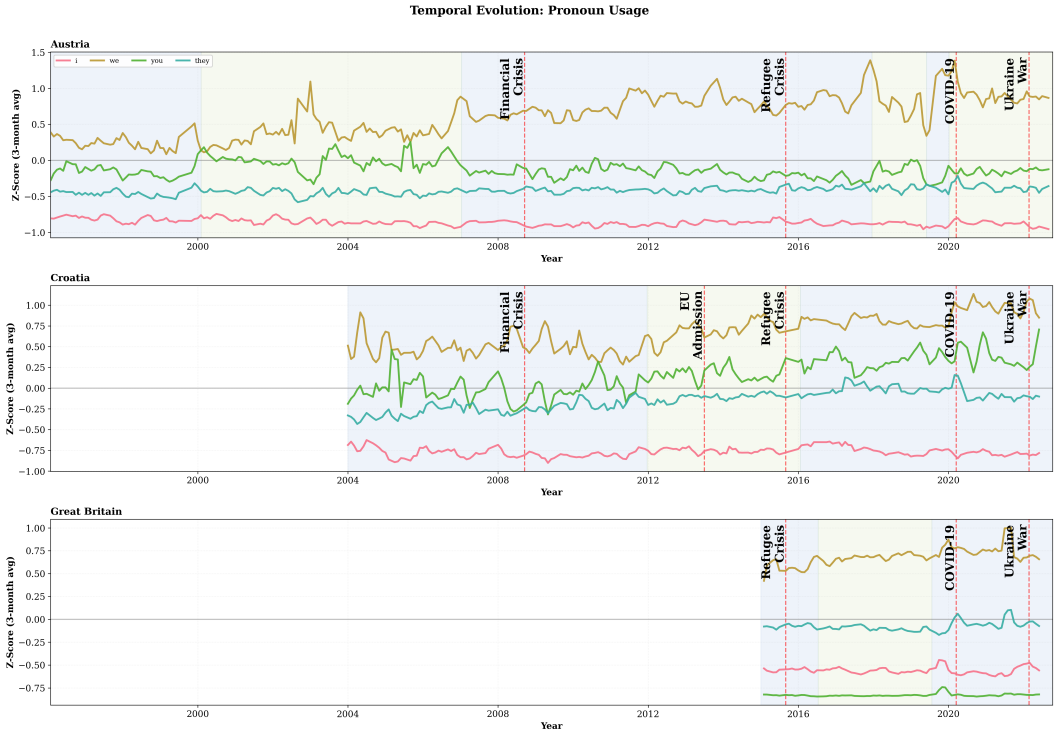


Figure 28.: **Pronoun use over time** (*I*, *We*, *You*, *They*). First-person singular remains consistently low, while collective `we` trends upward. Direct address `you` fluctuates with institutional routines such as Question Time.

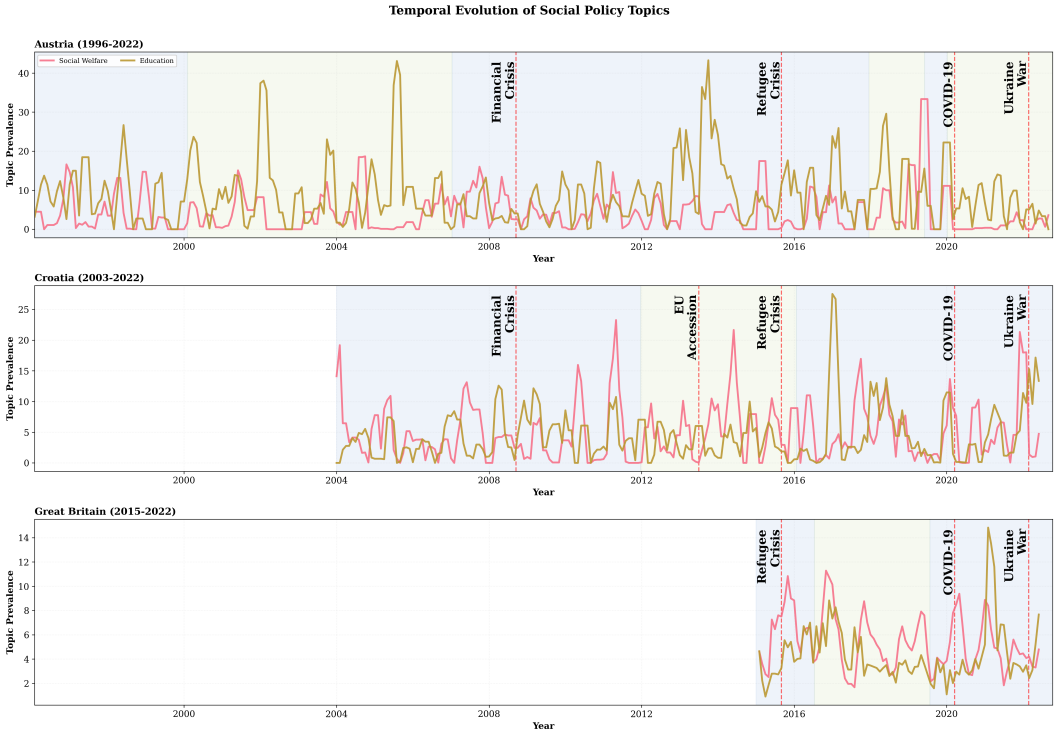


Figure 29.: **Social policy topics over time** (Social Welfare vs. Education). Social Welfare displays sharp spikes around major reform packages and fiscal negotiations, whereas Education shows smoother, longer-term trends.

D. Additional Figures from Section 4.4 (Austria)

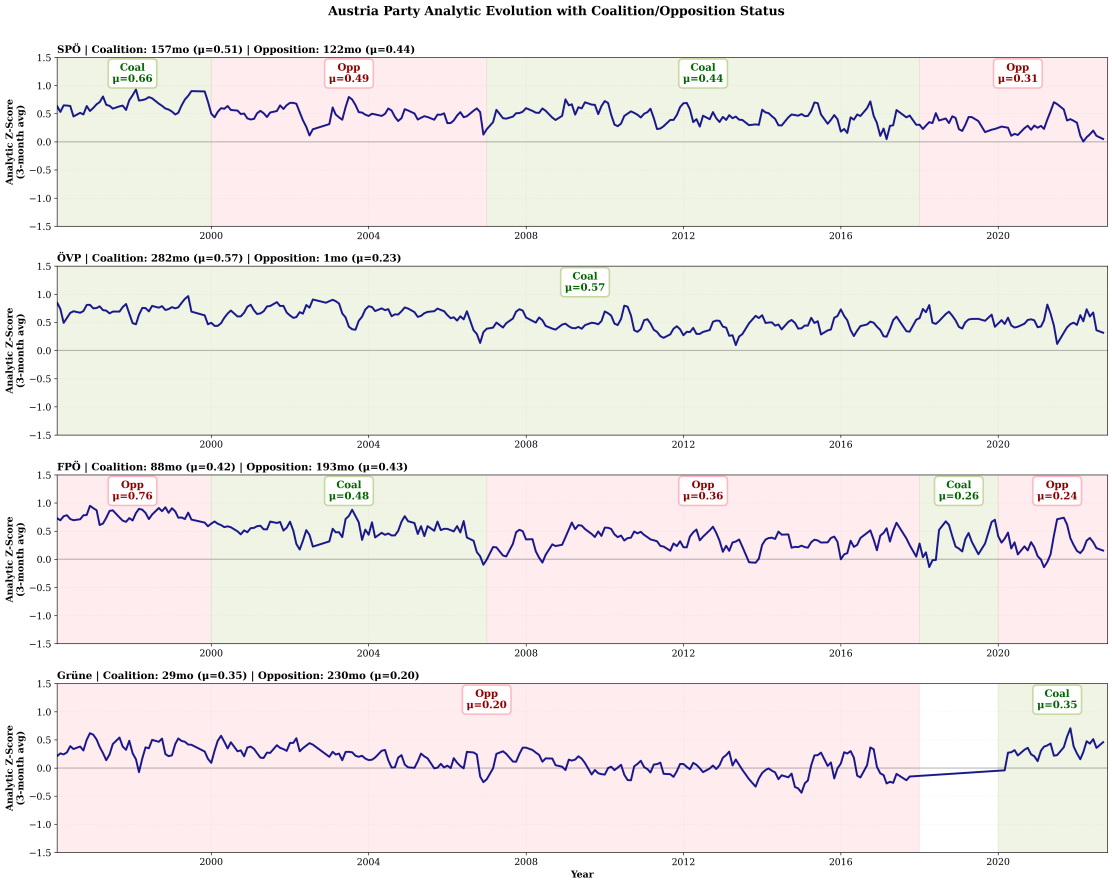


Figure 30.: **Austria: Analytic (party-level).** Three-month moving averages of *Analytic* by party. Governing parties tend to show slightly higher analytic scores, consistent with their responsibility for explaining and defending policy packages.

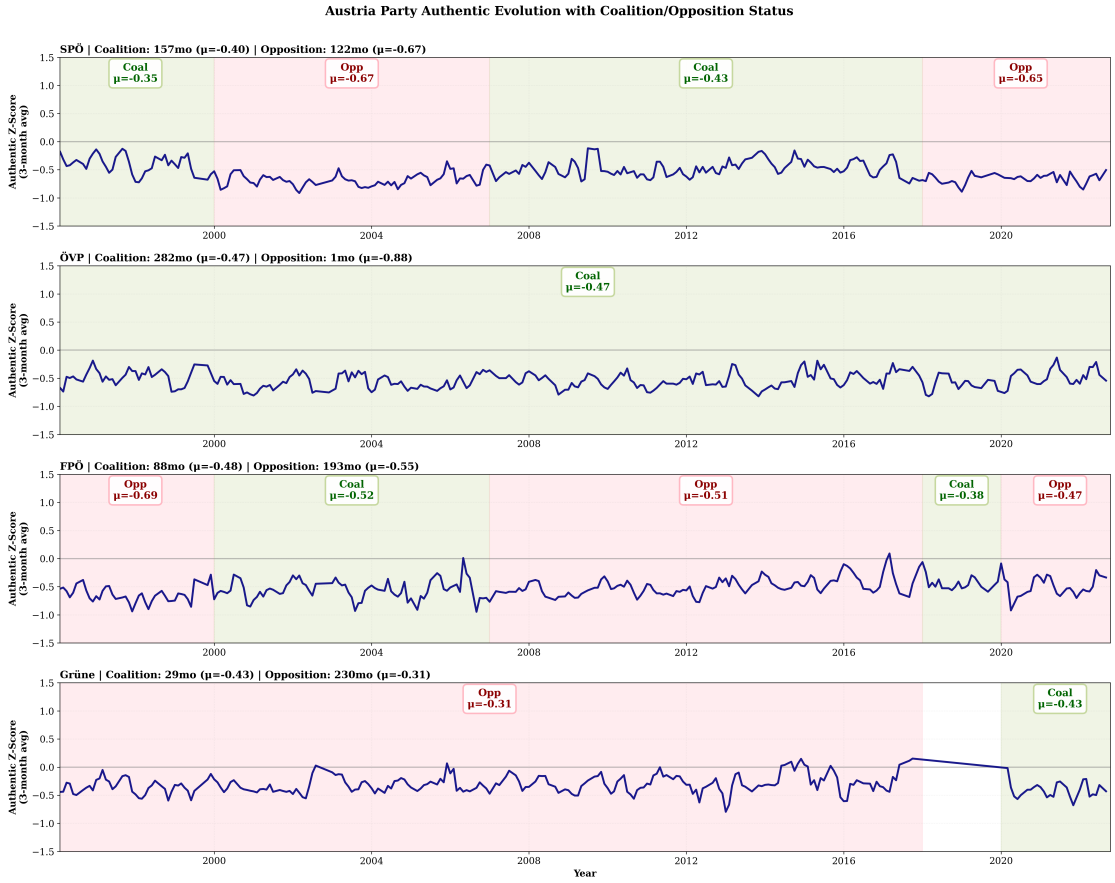


Figure 31.: **Austria: Authentic (party-level).** Authenticity fluctuates modestly across parties and time, with no simple alignment to government status.

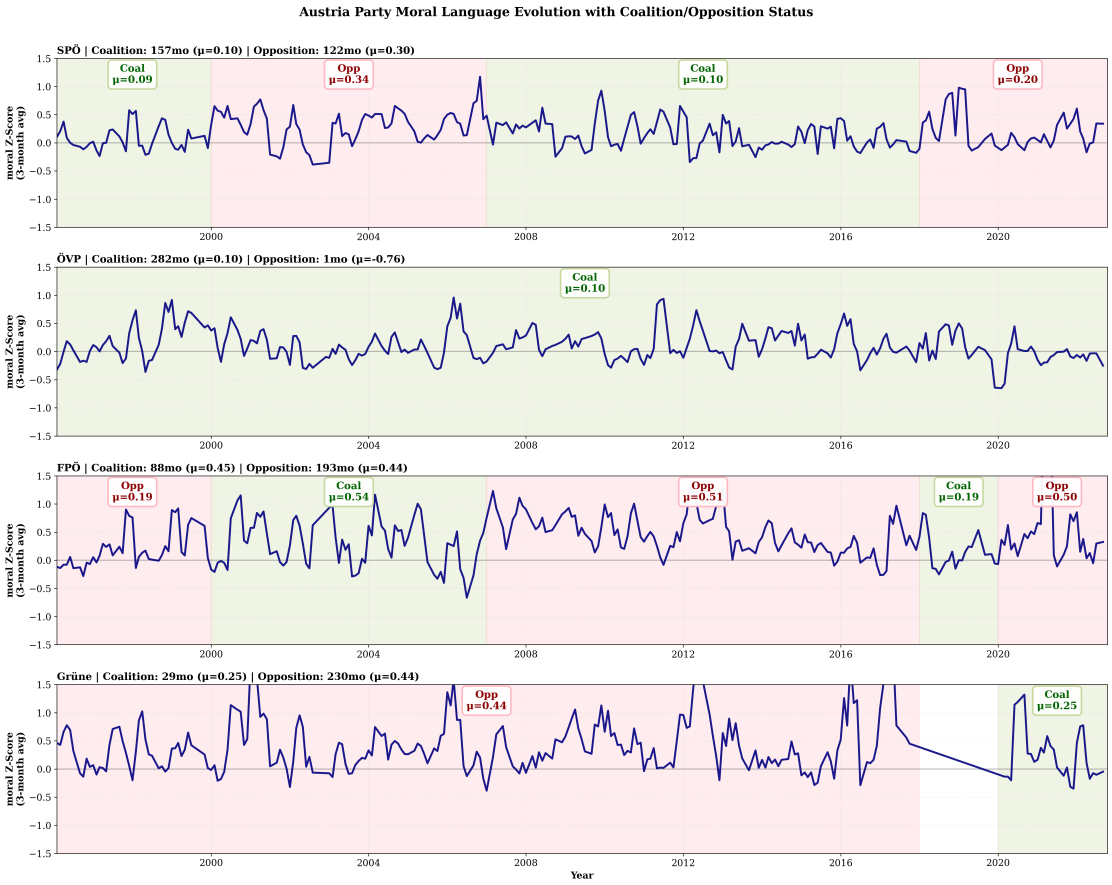


Figure 32.: **Austria: moral (party-level).** Moral language shows occasional spikes around debates on social issues and migration but remains relatively stable overall compared to political and power vocabulary.

Figure 33.: **Austria: anger (party-level).** Opposition parties display higher and more volatile anger scores, particularly during contentious legislative periods, while governing parties maintain lower, flatter profiles.

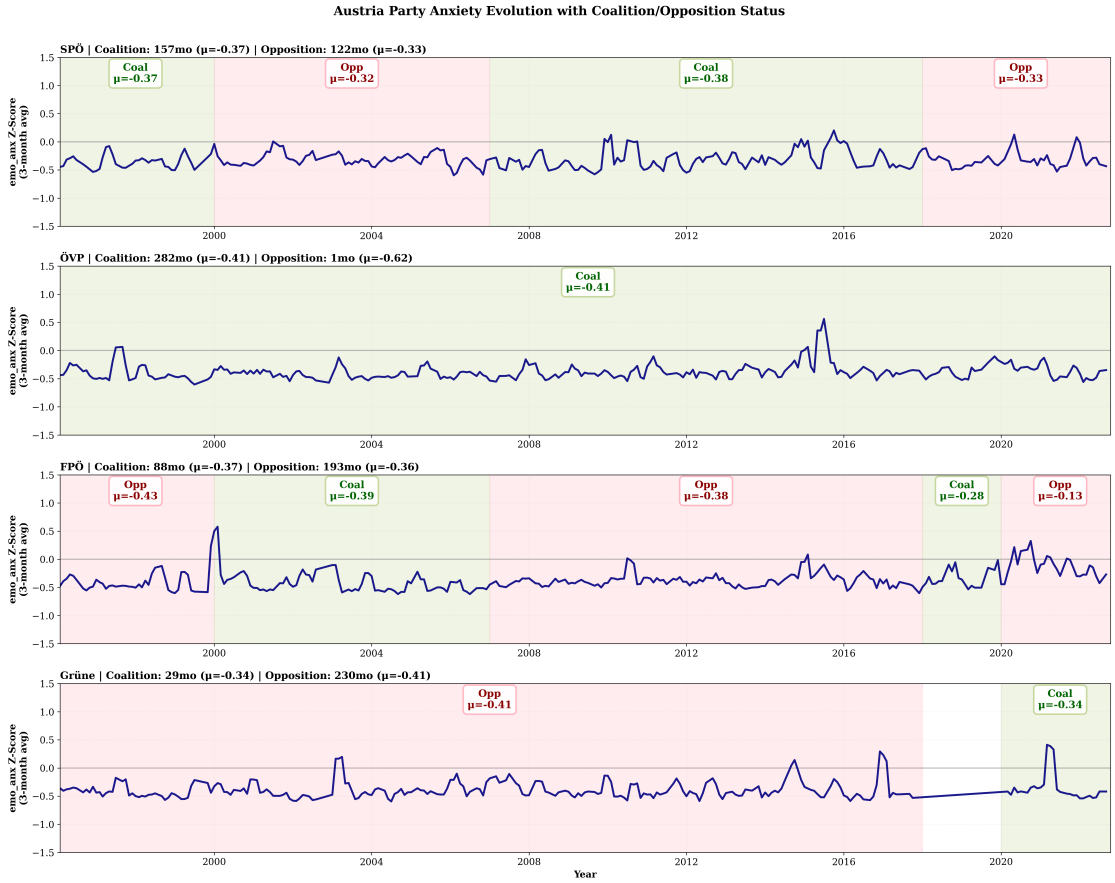


Figure 34.: **Austria: anxiety (party-level).** Anxiety-related language peaks around economic and migration crises and then recedes, with only small differences between parties once status is controlled for.

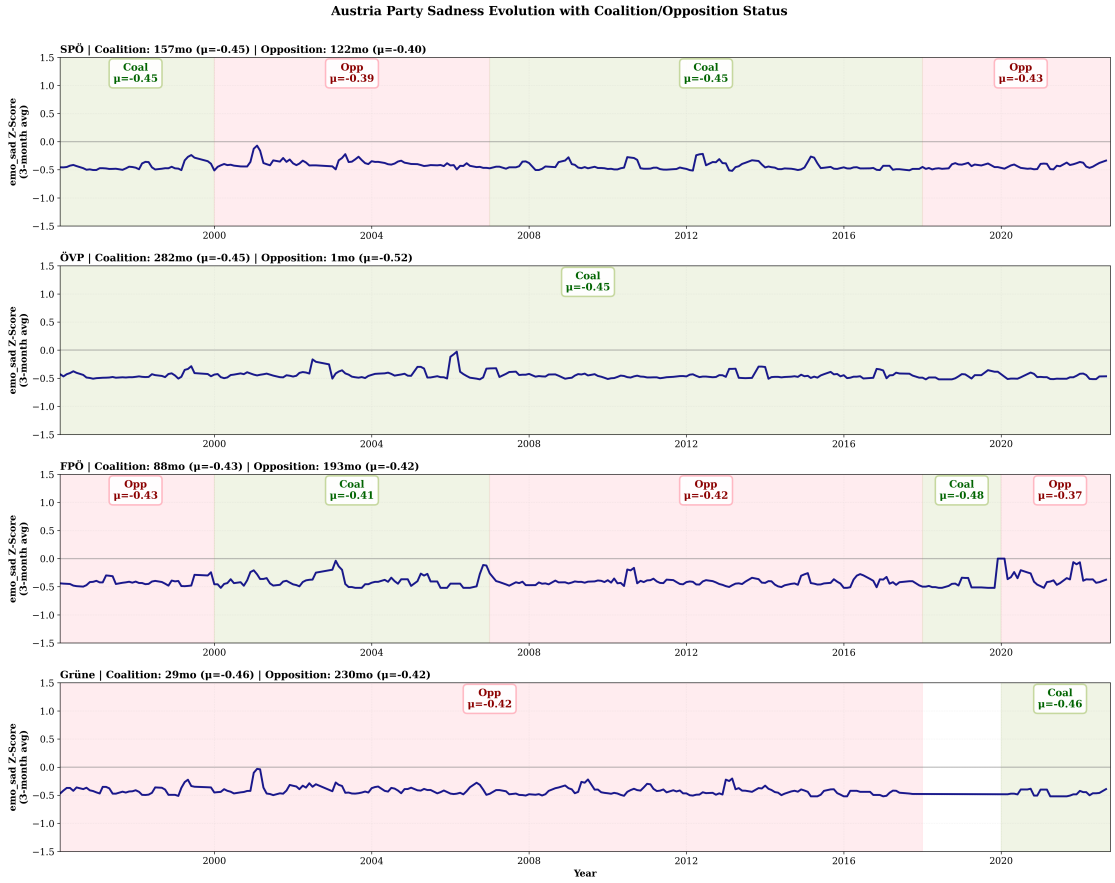


Figure 35.: **Austria: sadness (party-level).** Sadness increases during national tragedies and commemorative debates, affecting all parties similarly.

E. Additional Figures from Section 4.4 (Croatia)

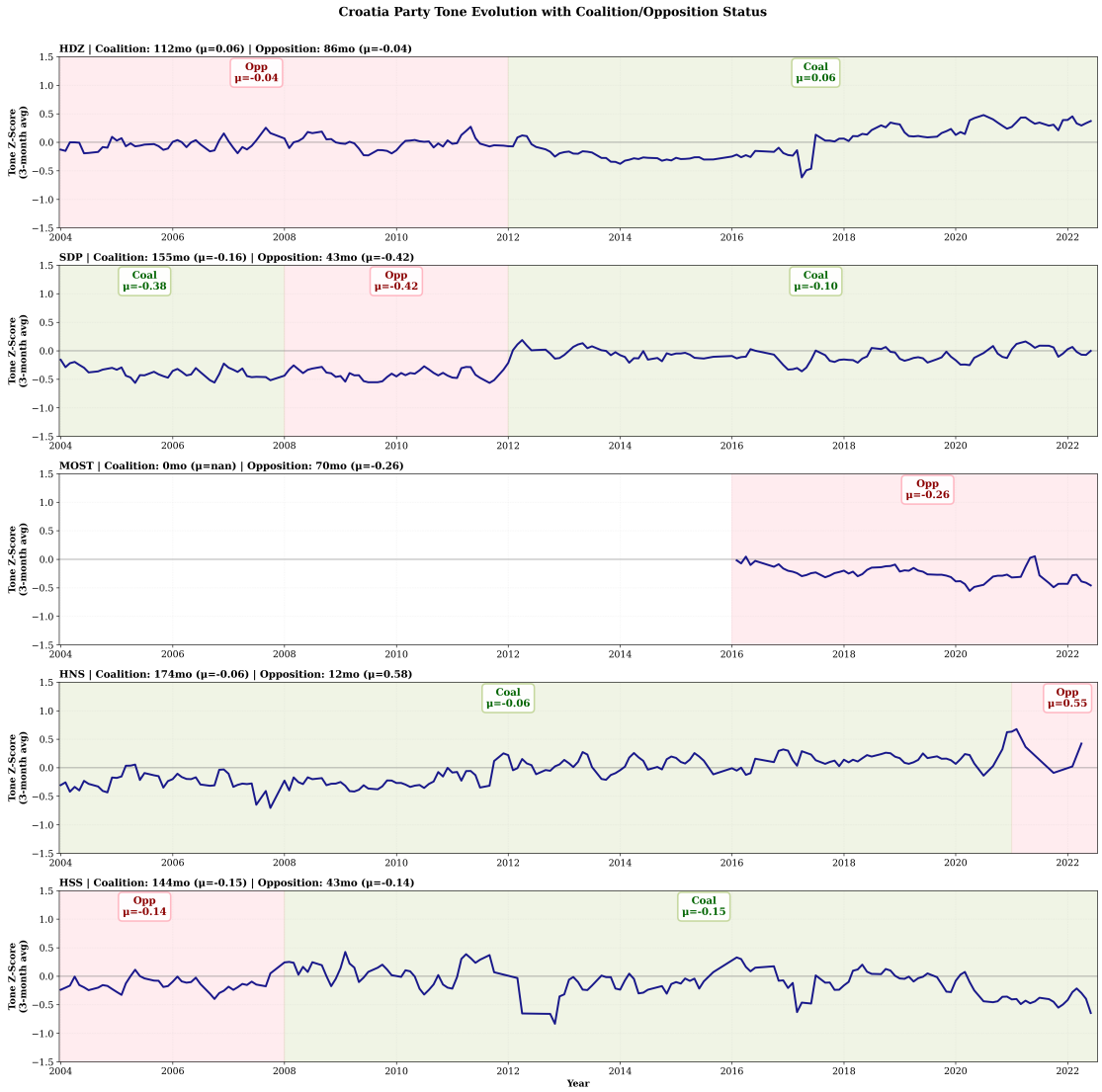


Figure 36.: **Croatia: Tone (party-level).** As in Austria, governing periods are associated with warmer language, while opposition periods show lower tone scores, especially during contentious reforms.

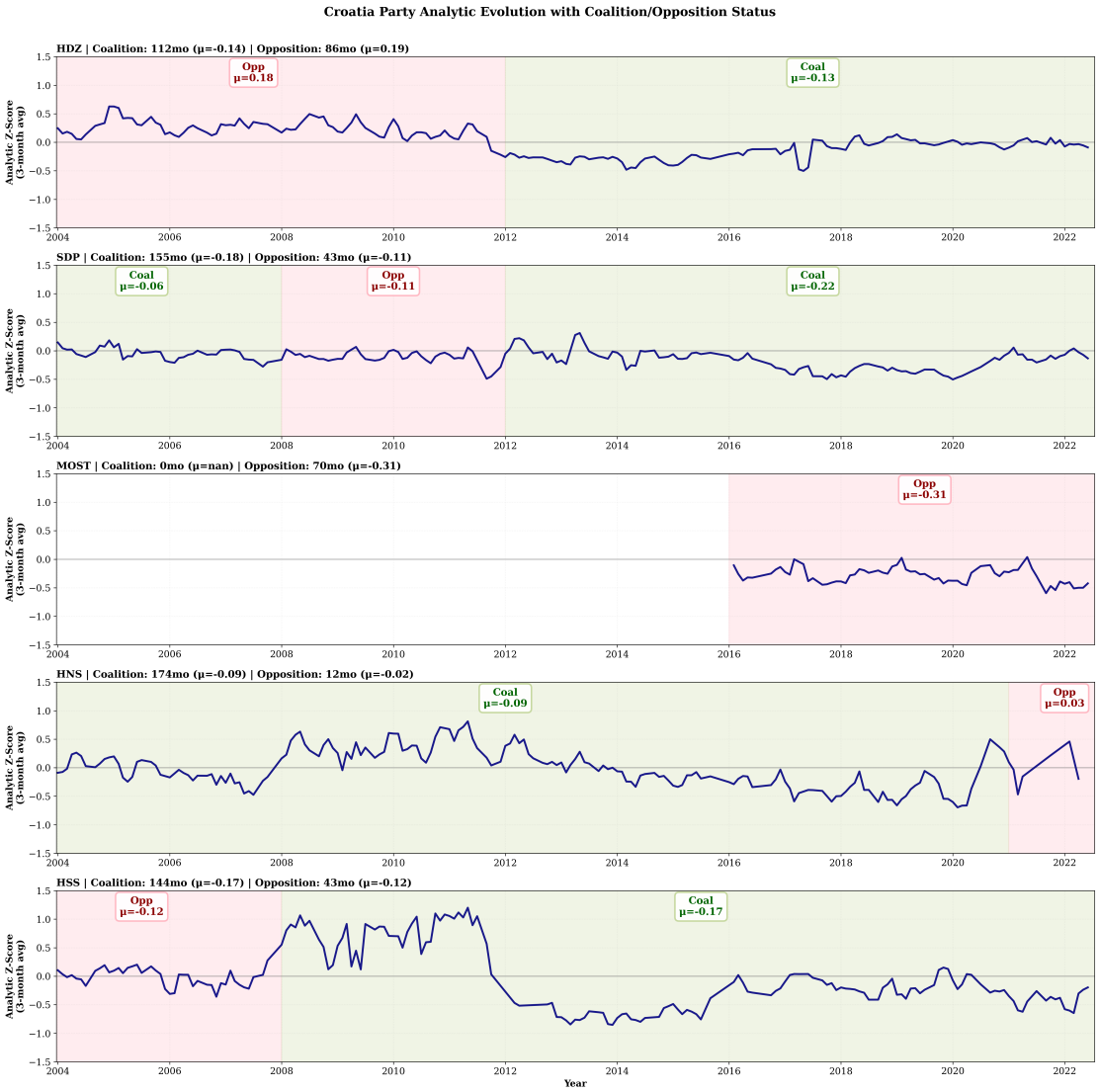


Figure 37.: **Croatia: Analytic (party-level).** Analytic style increases in reform-heavy periods and slightly more for governing parties, reflecting their role in presenting complex legislative packages.

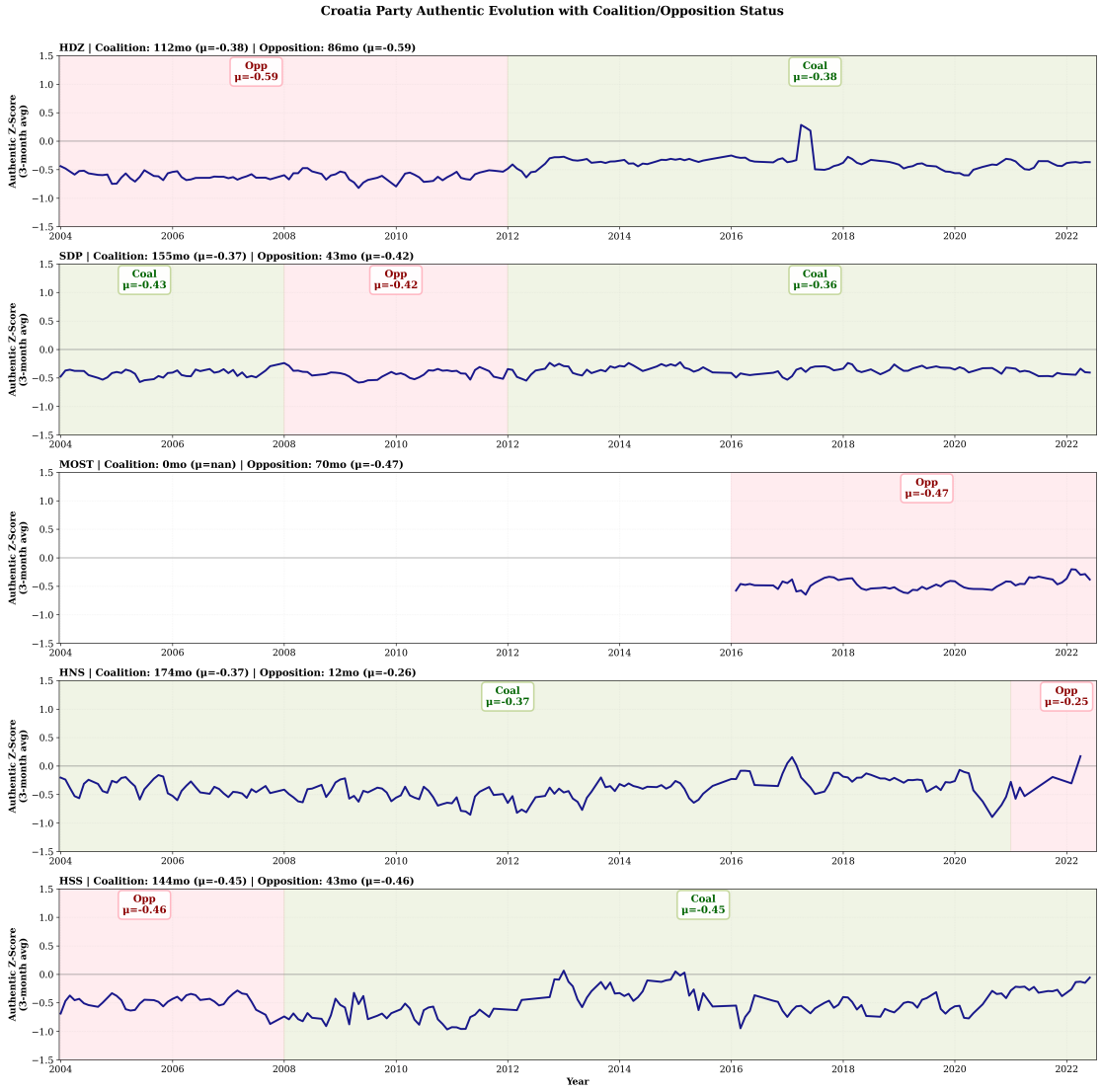


Figure 38.: **Croatia: Authentic (party-level).** Authenticity exhibits heterogeneous, party-specific dynamics, with some parties adopting a more personal and self-revealing tone over time.

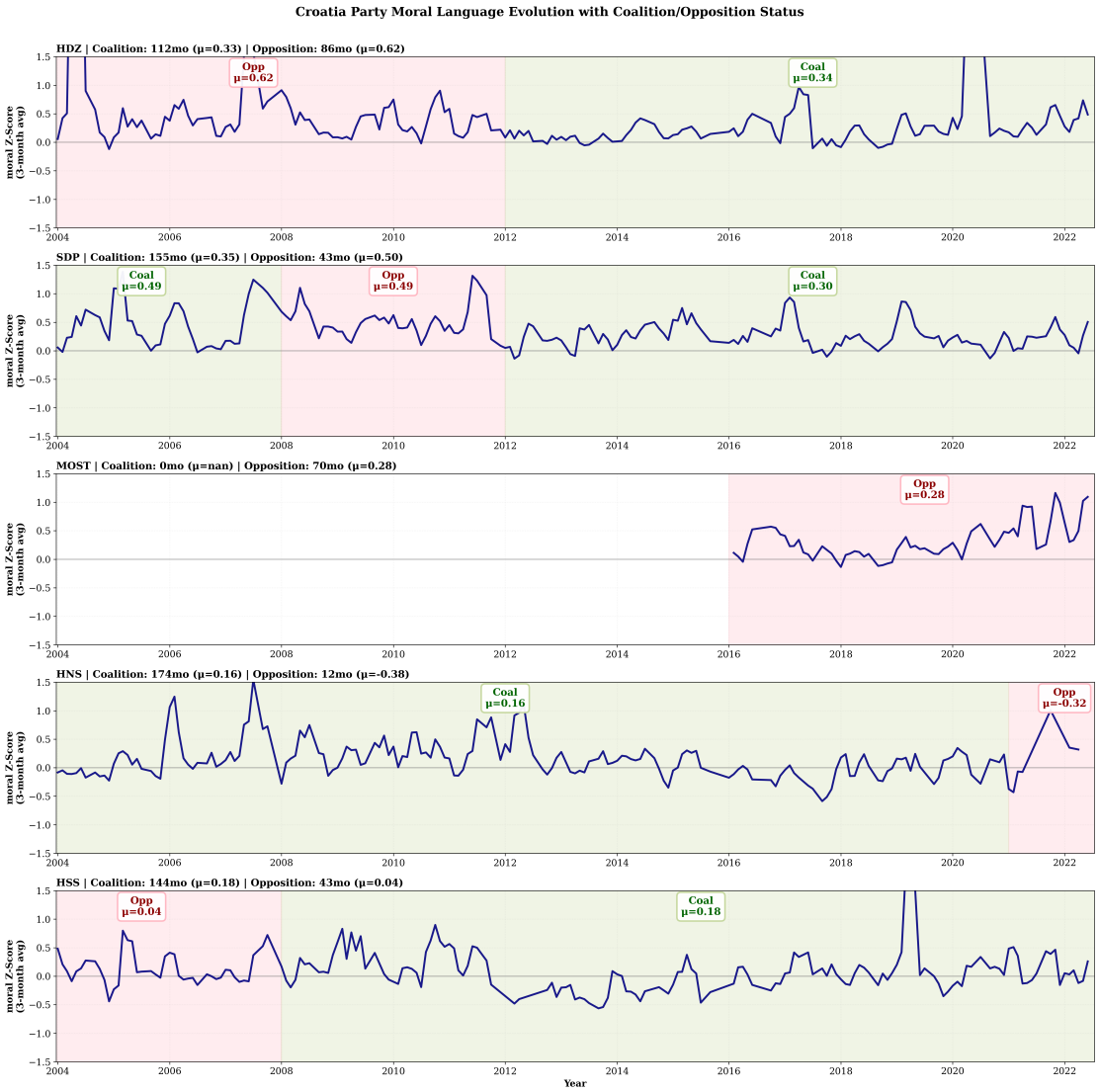


Figure 39.: **Croatia: moral (party-level).** Moral vocabulary intensifies around debates on family policy, education, and national identity, and is somewhat more pronounced in opposition speech.

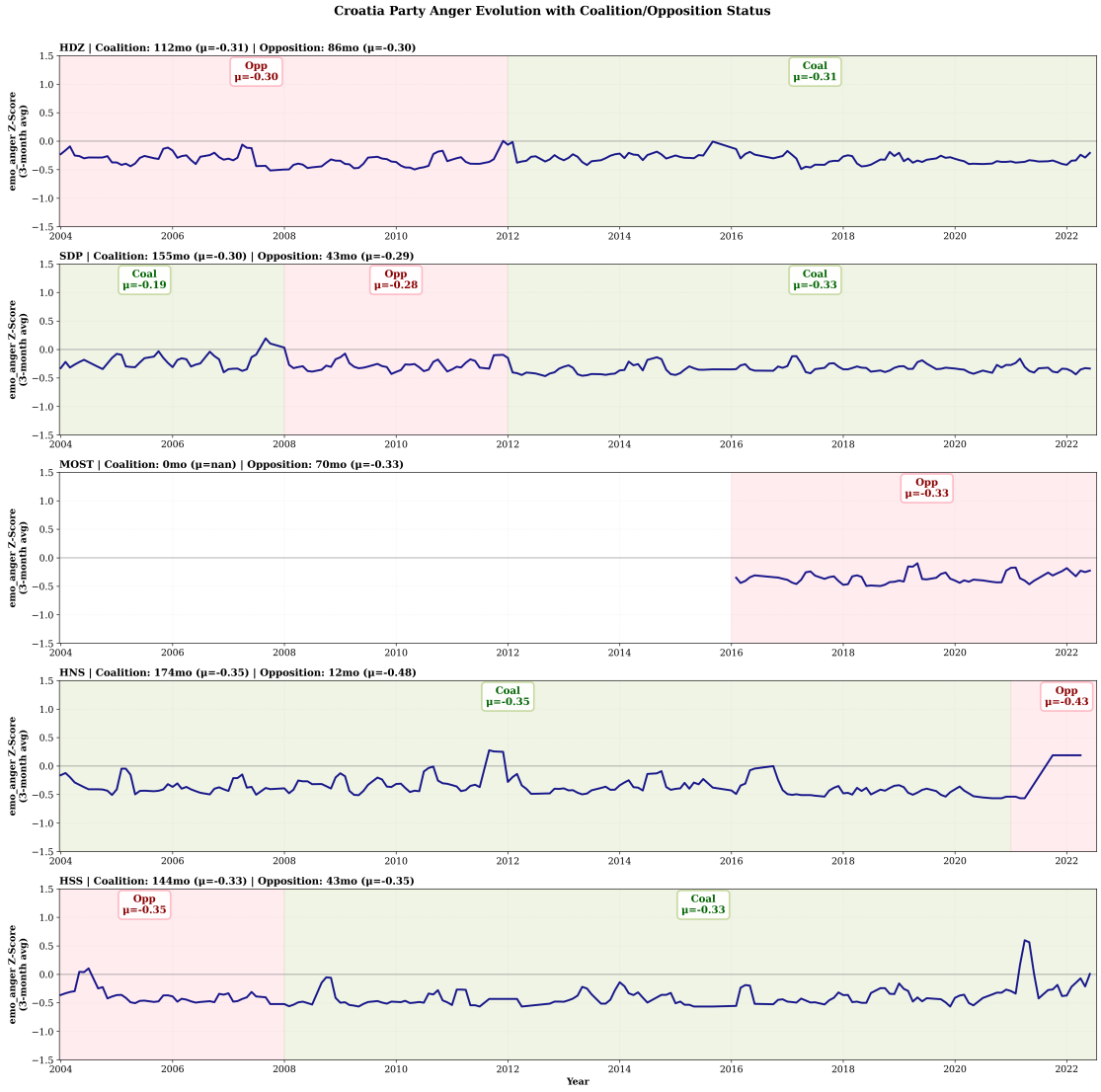


Figure 40.: **Croatia: anger (party-level).** Anger spikes in periods of political scandal and contested reforms, particularly within opposition parties, and then returns toward baseline.

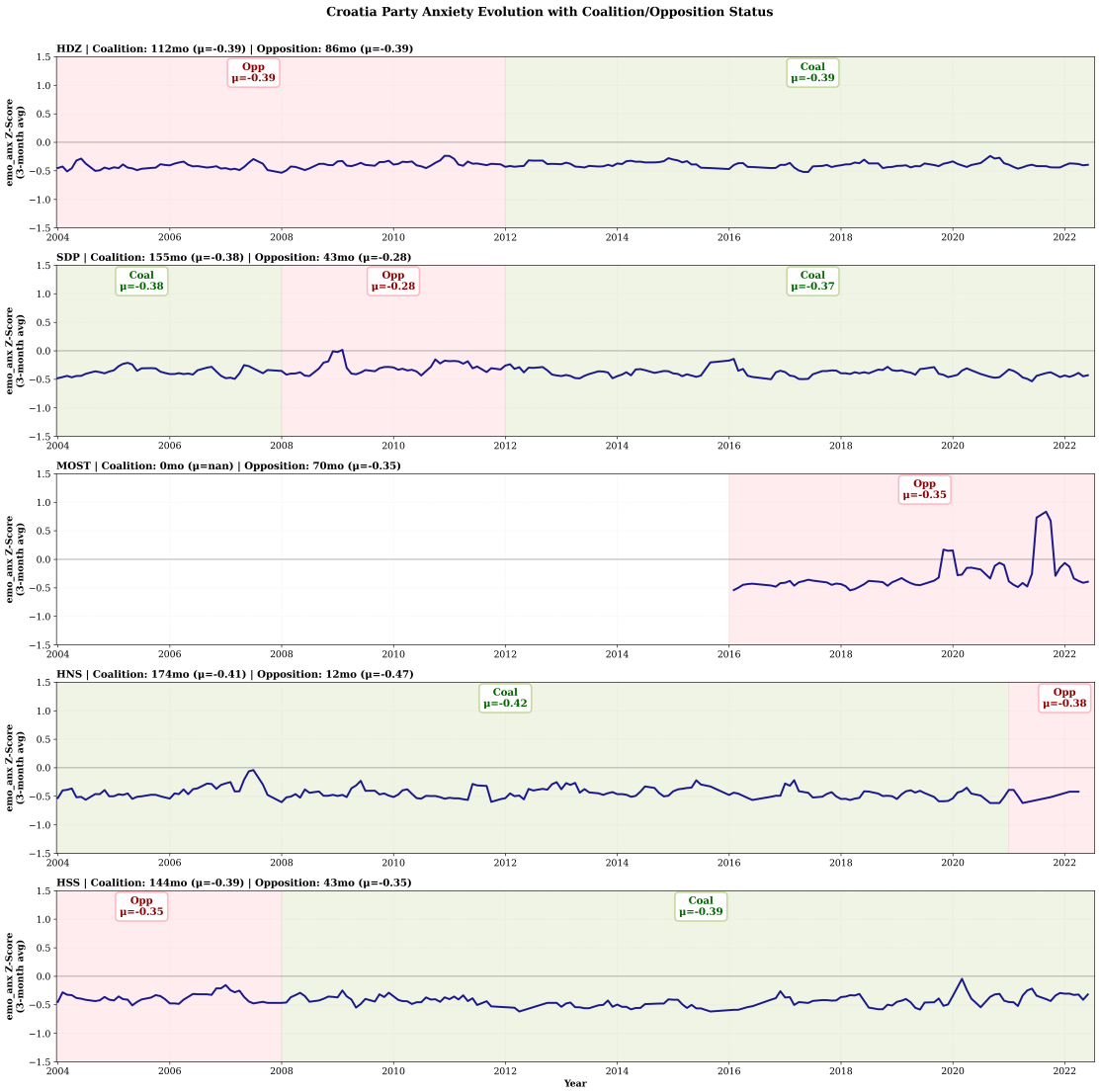


Figure 41.: **Croatia: anxiety (party-level).** Anxiety markers rise during economic and institutional crises and affect both government and opposition, though with somewhat higher levels in opposition parties.

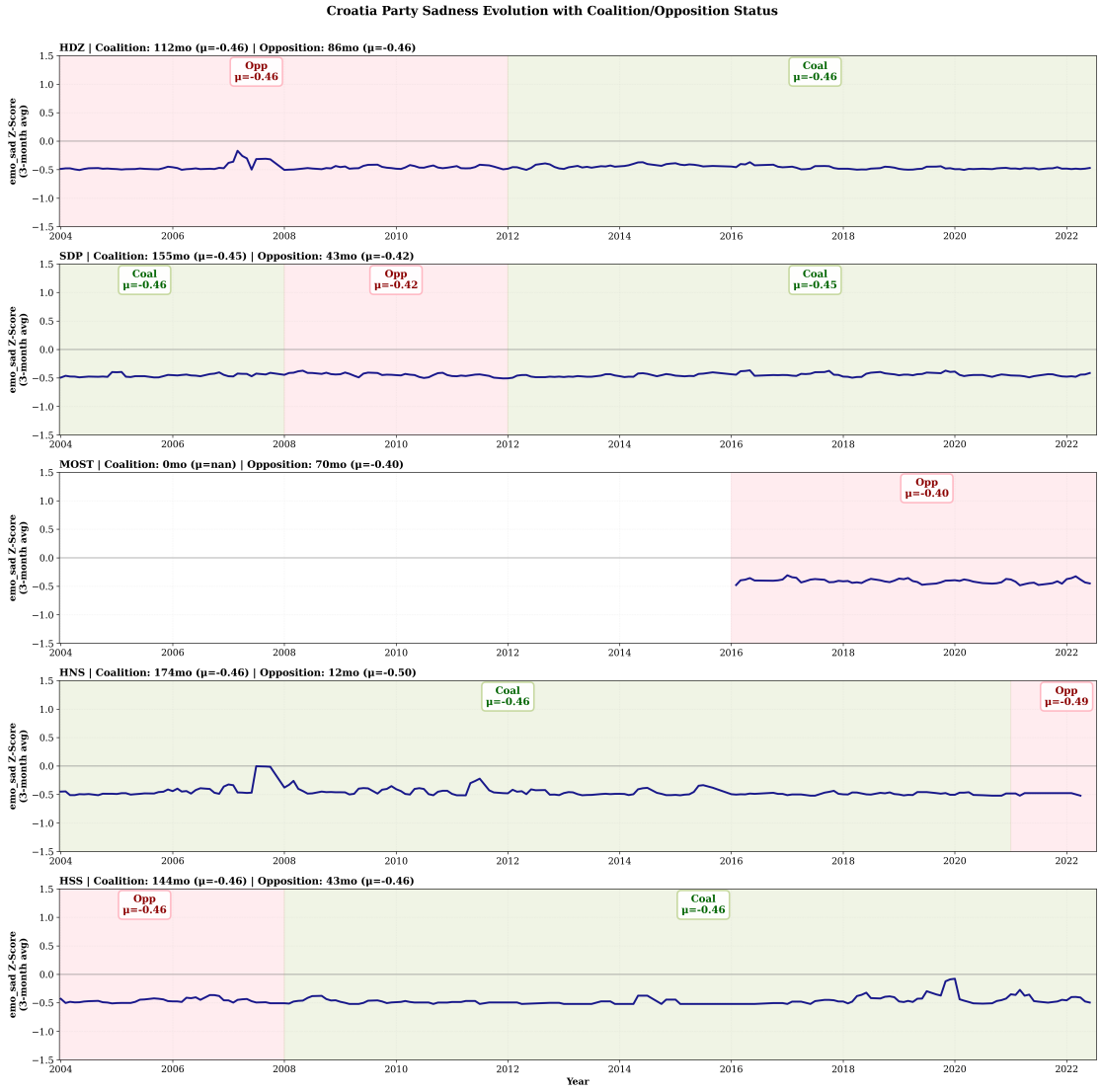


Figure 42.: **Croatia: sadness (party-level).** Sadness-related language increases during commemorations, natural disasters, and national mourning periods, cutting across party lines.