Amsterdam
University
Press

Back to the basics: Applying multilingual dictionary analysis to the Comparative Manifesto Project corpus

Joshua Cova
Max Planck Institute for the Study of Societies, Cologne

Abstract

For researchers interested in political communication and electoral politics, the Comparative Manifesto Project (CMP) is a widely-used database, which contains an extensive repository of annotated electoral manifestos. However, by relying on the fact that country-specific expert coders assign text excerpts to different policy areas, researchers frequently fail to engage with the multilingual nature of the corpus. This article uses a deductively-defined multilingual dictionary analysis to extract information on a set of electorally salient policies in seven European countries. In this application, I not only show that inter-coder reliability tests are encouragingly high and can be conducted relatively rapidly, but I also illustrate that the usage of multilingual policy-specific dictionary analysis is linked with high precision and high recall metrics. This approach can not only be helpful in examining electoral issue salience, but can also help in uncovering the language and the framing that political parties employ when discussing different policies.

Keywords: Comparative Manifesto Project, multilingual text analysis, dictionary analysis

Introduction

It seems safe to say that sometime during their careers, researchers who are interested in electoral politics will come across the corpus of the Comparative Manifesto Project (CMP) (Volkens et al., 2021). The CMP dataset offers data on electoral manifestos for over thousand parties from over fifty countries from 1945 onward and has, to date, been employed in almost six-hundred peer-reviewed academic publications.[1] Since its inception in 1979, the extensive CMP database has been leveraged by scholars interested in a variety of research questions and has proven to be an empirical testing ground against which to test foundational theories of party-political competition. In addition to providing an unrivaled repository of electoral

[1] https://manifesto-project.wzb.eu/publications/all

manifestos, the CMP provides multilingual *hand-annotated* party manifestos, which classify excerpts of text ('quasi-sentences' in CMP-speak) as belonging to different policy areas. This is useful for researchers, who are interested in conducting comparative studies in electoral politics. By relying on the prior annotation of the text of party manifestos, practitioners can avoid dealing with multilingual text analysis and can instead confide in the work conducted by expert coders in assigning sentences written in different languages to different policy areas. Nevertheless, as I show here, by depending on the prior annotation of electoral manifesto data, analysts might be constrained in their choice of how to analyze the way in which political parties discuss different policies, which, in turn, has consequences for the substantive interpretation of their findings.

This article begins by providing an overview of the usage of the CMP corpus as a data source, its use in political science research and the potential issues which analysts confront when using CMP data for comparative analyses. This is followed by an illustration of some prior work on multilingual text analysis and an overview of how insights from this area of research can help address methodological problems pertaining to the specific nature of the CMP dataset. Compared to the voluminous research output, which the usage of modern computational techniques has given rise to in different monolingual settings, research in multilingual computer-assisted content analysis is still a relatively new field of research (Lind et al., 2019; Lucas et al., 2015).

In this article, I illustrate how a multilingual quantitative text analysis, which uses mostly deductively-constructed dictionaries, validated by native speakers, can allow analysts to focus on policies as opposed to pre-assigned and CMP-defined policy areas. In order to illustrate the advantages, which can derive from the usage of multilingual dictionary analysis, this article conducts a multilingual dictionary analysis on the corpus of the CMP, by examining three electorally salient as well as divisive policies (minimum wages, same-sex marriage and regulations on the termination of pregnancy) across seven European democracies (Austria, Belgium, France, Germany, Ireland, Switzerland, United Kingdom) in three different languages (French, German, English). In this application, I focus on illustrating the salience which these policies have in electoral manifestos as well as the broader semantic context in which these are articulated. I show that the automated application of multilingual dictionary analysis constitutes a valid and efficient alternative to the manual recoding of quasi-sentences as well as the more generic 'shortcut' of using CMP codes to avoid dealing with the multilingual complexity of the CMP corpus. Validation checks not only illustrate that inter-coder reliability metrics yield satisfactory results, but that this approach also results into high levels of recall and precision; two frequently employed performance metrics in the specialized literature on data and information retrieval. Moreover, by focusing on policies, which can be identified through custom-built dictionaries, researchers can exercise greater control on which issues they are interested in analyzing and are not constrained by the way in which manifesto coders categorize text.

Quasi-sentences, multilingual text analysis, and the limits of the Comparative Manifesto Project's policy areas

Electoral manifestos constitute an important source of information, which is used by the media, citizens and researchers in order to discern political parties' positions on different policies and topics. Electoral manifestos provide an indication of which policies, parties are likely to implement when and if in power, and constitute a guide to political parties' policy priorities (Allen & Bara, 2019; Klingemann et al., 1994). As a consequence, the CMP, which rests on the extensive coordination efforts of researchers, specialized in the politics of different countries, has become the 'canonical' dataset used by researchers of comparative electoral politics (Benoit et al., 2016).

The fundamental unit of analysis for CMP data is that of the quasi-sentence, which in the CMP's coding instructions is defined as a part of text which contains 'one statement or "message"'.[2] As the length of natural sentences varies between party families, languages and countries, quasi-sentences are designed to capture the core parts of the sentence, which denote the policy message a party wishes to convey. The fifty-six policy areas in which quasi-sentences can be categorized in are defined deductively and are grouped into seven broader domains (1. External relations, 2. Freedom and democracy, 3. Political system, 4. Economy, 5. Welfare and quality of life, 6. Fabric of society, 7. Social groups). These predefined categories seek to encompass the wide range of policy areas political parties can be expected to compete on.

From a linguistic perspective, it is important to note that while the CMP coding guidelines are 'language-independent', coders apply their language-specific knowledge to the manifesto in order to divide natural sentences

[2] Manifesto Coding Instructions: 5th fully revised edition (2021), p.5

into quasi-sentences and subsequently classify quasi-sentences into their appropriate categories. This is therefore something that end-users have no control on. Moreover, CMP coders are instructed to classify quasi-sentences based on the policy goals that these excerpts convey; as stated in the coding instructions: 'goals usually take precedence over means when assigning codes'[3]. This entails that the CMP coding scheme does not necessarily concentrate on the policies which are articulated in the manifestos, but rather seeks to focus on the objectives present in electoral platforms.

The fact that quasi-sentences are categorized into different policy areas has been leveraged by researchers interested in examining the question of how frequently parties discuss different policy areas in their electoral manifestos. The underlying assumption of this approach is that the higher the number of quasi-sentences dealing with a specific policy area, the more *salient* is a policy area expected to be for political parties (Budge & Farlie, 1983; Petrocik, 1996). Nevertheless, several researchers have expressed concerns about the way that quasi-sentences are coded. While some critical work has argued that some quasi-sentences seem to be misclassified (Zulianello, 2014), other researchers have focused on the fact that the results of inter-coder reliability tests, designed to test whether different coders agree on the classification of quasi-sentences, have been underwhelming (Däubler et al., 2012; Mikhaylov et al., 2012). While these are important criticisms that researchers using the CMP corpus should acknowledge, this article addresses another issue pertaining to the usage of CMP-assigned quasi-sentences. Although quasi-sentences are designed to tease out key political messages from longer natural sentences, the policy categories identified by the CMP tend, typically, to be too broad to capture the specific policy areas that analysts are often interested in. Thus, in their analysis of how parties compete on ethno-cultural themes, Protsyk and Garaz (2013) argue that the CMP's coding scheme 'lacks adequate granularity to capture the subtleties of party rhetoric in the particular policy area' (p.297). In other words, the policy area codes offered by the CMP on this topic are too broad to capture the precise policies researchers are interested in. The solution adopted by Protsyk and Garaz (2013) is to code manually the relevant text excerpts, which were identified based on their CMP codes, and to then define a more specific coding scheme. Similar manual recoding efforts by Horn et al. (2017) have, based on the relevant CMP code, sought to further understand the way in which German parties discuss welfare state expan-

[3] Manifesto Coding Instructions: 5th fully revised edition (2021), p.9

sion in their electoral manifestos. Farstad (2018) has instead examined how climate change is portrayed in West European party manifestos by selecting quasi-sentences belonging to the more general CMP code ’environmental protection’, where, she argues, policy statements on climate change are most likely to appear.

The studies presented above have thus focused on re-coding quasi-sentences by incorporating greater granularity into the coding scheme. While this approach allows for a greater refinement and, presumably, confers a greater validity to text analyses, it is also problematic for three reasons. First, this is a labour-intensive approach, which in addition to designing and validating a novel coding scheme also requires parsing through all the relevant quasi-sentences of the party manifestos researchers are interested in. Second, this approach assumes - often implicitly - that policies are appropriately categorized within the correct policy area; for instance, it assumes that policy statements on climate change are only coded in the subset of quasi-sentences coded in the CMP policy area ’environmental protection’ and *do not* appear in quasi-sentences coded as belonging to different policy areas. As my analysis will illustrate this assumption is often unfounded. Third, users who re-code quasi-sentences do, by definition, utilize the quasi-sentence as their unit of analysis. However, considering that quasi-sentences describe an objective, a message, as defined by the CMP coding scheme, and do therefore not necessarily focus on a policy or an issue, this might not constitute the best strategy for researchers interested in pursuing a policy-based analysis of electoral manifestos.

Changing policy priorities and new party constellations

In this article, I consider the usefulness of applying a multilingual dictionary analysis on electoral manifestos by examining a set of countries (Austria, Belgium, France, Germany, Ireland, Switzerland, United Kingdom) and policies (minimum wage, abortion, same-sex marriage) in the timeframe 2000-2021. The decision to study these countries and policies is motivated by the fact that these policies have in recent times become increasingly politicized. In the past decades, for example, same-sex marriage has been a hotly debated topic, which has been legalized only recently in the countries that are within this study’s purview: Austria (2019), Germany (2017), France (2013), Switzerland (2022). More generally, in light of the recent electoral successes registered by right-wing populist parties (RWPPs), abortion rights, same-sex marriage and other issues related to morality politics have become increasingly electorally contested.

The relationship between morality and gender politics and RWPPs has been subject to competing assessments in the academic literature. While on the one hand, RWPPs have, for instance, been seen as a quintessentially traditionalist force, with traditionalist views on issues, such as family composition and reproductive rights, RWPPs such as Geert Wilders' Party of Freedom in the Netherlands have also come out in favour of progressive policies on civil and sexual rights, which are then counterpoised to less liberal values purportedly espoused by migrant communities (Akkerman, 2015; Dietze & Roth, 2020). As there seems to be uncertainty on the way in which radical right parties position themselves on morality politics, some scholars have found it more fruitful to focus on the cross-country differences, which influence the salience of morality politics in the party political discourse. This for example constitutes the premise on the influential work conducted by Engeli et al. (2012) on the 'two worlds' of morality politics; in which it is argued that the past presence of a clear historical conflict between religious and secular parties shapes the salience of morality politics as an electoral issue in the present era.

Here I also examine a broadly popular economic policy, that is the minimum wage, whose political salience has increased in different European countries over the past decades. The UK, Ireland and Germany introduced a statutory minimum wage relatively recently (2000, 2001 and 2015 respectively). While discussions on whether to introduce a statutory minimum wage in Austria and Switzerland are ongoing (Schulten, 2014), minimum wages have become increasingly politicized policies across different European countries (Cova, 2023). Considering that minimum wages have historically been a priority for left-wing parties, one would expect that left-wing parties should discuss minimum wages in electoral manifestos more than their right-wing counterparts. The relationship between partisanship and the salience of minimum wages might however be more complex than this overview would suggest. Indeed, it is important to differentiate between the parties that are collocated to the left of the political spectrum, which have placed greater emphasis on socio-economic issues (e.g. radical left parties) versus post-materialist issues (e.g. Green parties) (Hooghe et al., 2002). Expectations on right-wing parties are also varied. While economically liberal, right-wing parties, might be expected to assign a lower salience to minimum wages, this expectation might need to be revised in the case of Christian-Democrats, which have tended to favor greater state intervention in the market. Right-wing populist parties have instead often been found to prefer consumption-led policies favoring low-wage workers (Chueri, 2022); this might entail that the salience of minimum wages in RWPPs electoral manifestos might be higher than for other right-wing parties.

To summarize, the choice of countries and policies is not only motivated by the need of including different languages for the purposes of the multilingual analysis, but has also been informed by the preference of selecting a set of politically charged topics, which are clearly relevant for the countries examined here. While this article seeks to make a methodological-focused contribution, it nonetheless hopes to provide insights on the way in which different party families across different countries have discussed these politicized issues in the past decades in their electoral manifestos.

Dictionary analyses applied to a multilingual corpus

This article considers the option of analyzing through the use of custom-built and multilingual ’policy dictionaries’ policies as opposed to CMP-defined policy areas. Compared to the more sophisticated approaches offered by natural language processing techniques, dictionary analyses rest on a fairly simple and intuitive approach. Analysts define, usually deductively, a list of keywords, which is of interest to them and subsequently examine the frequency with which these keywords occur in a corpus. Nevertheless, also quantitative text analyses which rest on a dictionary approach can vary in complexity. Amongst the simplest quantitative text analysis approaches, which make use of dictionaries, is that of visibility analyses, which examine the frequency with which certain persons/actors are mentioned in the media (Rubin et al., 2021; Vos & Van Aelst, 2018). In addition to their simplicity, visibility analyses have high validity; provided that the names (the keywords, in this case) are spelled correctly and are ’exhaustive and correct’ (Boumans & Trilling, 2016), a computer-assisted visibility analysis should be able to recall all instances in which the persons of interest are mentioned in the corpus. While visibility analyses have high validity, they also tend to have low generalizibility, as they rest on custom-built keyword lists, which probably cannot be applied to other contexts.

Dictionary analyses do, however, become more complex once researchers

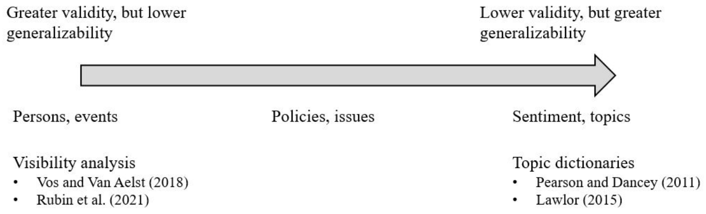
Greater validity, but lower generalizability
Lower validity, but greater generalizability
Persons, events
Policies, issues
Sentiment, topics
Visibility analysis
• Vos and Van Aelst (2018)
• Rubin et al. (2021)
Topic dictionaries
• Pearson and Dancey (2011)
• Lawlor (2015)

Figure 1: Different types of dictionary analysis

focus on examining more generalizable, but less specific topics. This is the case for topic dictionaries, which list a set of keywords which can be linked to a certain topic. Topic dictionaries have frequently been employed when examining how different newspapers or political parties frame discussions on different issues (Albaugh et al., 2013; Lawlor & Tolley, 2017). In an analysis of how the British and Canadian print media discusses immigration, Lawlor (2015) draws on monolingual topic dictionaries to create a set of ’framing dictionaries’ (e.g. a crime frame dictionary, an unemployment frame dictionary), which are then used to classify news items discussing the topic of immigration from different perspectives. However, when using topic dictionaries, analysts need to be mindful of two problems. First, topic dictionaries might not be exhaustive (that is they are incomplete) and might, as a consequence, miss keywords that could be of interest. Second, keyword lists might not be mutually exclusive and could realistically be included into different domains. As an example, Pearson and Dancey (2011) investigate what gender differences can be discerned when examining US Congress members’ speeches on ’women’s issues’ legislation. To examine this question, the researchers construct a custom-built dictionary, with a list of keywords they identify as relating to ’women’s issues’ (e.g. woman, girl, female, servicewoman). While this approach confers high recall to the results (that is, it is unlikely that there will be speeches discussing legislation on ’women’s issues’ without using the keywords that were identified by the authors), it could also be that policymakers use the word ’woman’ in passing, in a context that is unrelated to legislation that could be classified as pertaining to women’s issues. This means that this approach might suffer from precision issues. For custom-built, context-specific topic dictionaries, validation is therefore extremely important (Grimmer & Stewart, 2013); even more so in a multilingual setting (Lind et al., 2019).

In this article, I examine whether by focusing on policies, one can circumvent some of the classification problems related to dictionary analyses. Compared to the broader CMP policy areas, policies deal with very specific issues. Same-sex marriage and minimum wages are, for example, two issues, with few synonyms and alternate signifiers. While text excerpts from electoral manifestos might refer to civil partnerships within the context of discussions on same-sex marriage, the term ’same-sex marriage’ is conceptually different from a civil partnership. Moreover, compared to the example presented above on the representation of ’women’s issues’ in the US Congress, it is unlikely that specific keywords associated with certain policies will refer to something else. Signifiers denoting a certain policy are unlikely to carry over into other policy domains. In other words, it is improbable that electoral manifestos would, for example, employ the words ’termination of pregnancy’, while discussing policies that are unrelated to discussions of abortion rights (high precision, but low recall). Therefore, while some words which are assigned to a specific topic dictionary can be conceivably collocated in other frames as well, this is less likely to be the case for policy dictionaries, which only examine specific issues. It thus seems reasonable to expect that this approach results into high recall figures: something for which, as the next section shows, I find empirical evidence for. Moreover, as I illustrate, these findings are robust across a multilingual dictionary analysis specification as well.

In this article I examine the way in which parties competing in different countries and in different languages discuss policies in their electoral manifestos. This is why the first step I take is that of constructing a multilingual policy-specific dictionary. Several scholars from different fields of research have noted (and lamented) the paucity of quantitative text analyses employing a multilingual lens of analysis (Dashtipour et al., 2016; Lucas et al., 2015; Proksch et al., 2019). Lind et al. (2019) in their guidance of how to construct a multilingual dictionary suggest the following workflow: 1. Keyword pre-selection, 2. Keyword translation, 3. Keyword evaluation. As a first step, I thus deductively codify (pre-select) a list of keywords, which can be associated with the policies I am interested in examining. As shown in the appendix, the list of keywords refer to synonyms or words that are strongly connected with the respective policies. Although the resulting multilingual policy-specific dictionaries are small in size they can, as it will be shown later on in the text, lead to satisfactory performance metrics. While to the best of my knowledge, an analysis similar to the one presented in this article has not been conducted thus far, as a point of comparison, the multilingual keyword lists used here are minuscule compared to those of topic dictionaries. For instance, Lind et al. (2019) use multilingual topic dictionaries, containing almost 200 keywords per topic, in their analysis of the performance of multilingual framing dictionaries in the classification of news articles discussing immigration. Policy-specific dictionaries are likely to be considerably smaller. In this article, they range in size from 2 to 12 keywords per policy.

In order to identify which keywords can be associated with different policies, I employ a deductive approach that is informed by the guiding principle that the policy-specific keyword list needs to be as specific as possible and pertain only to the policies, which are within this article's scope. The selection of keywords therefore seeks to minimize the risk of collecting generic keywords which can refer to other policy domains, while at the same time it seeks to include as many keywords, which can *exclusively* be associated to the topic of interest. While the construction of dictionaries typically combines deductive as well as inductive approaches (Baden & Stalpouskaya, 2015; Rauh, 2018), the specificity of the keywords in this context is so high that only one term has been added inductively as a result of the validation analysis, which will be explained below. In constructing the multilingual policy-specific dictionaries it is important to not only focus on the language, but also on the country-specific contexts, which predicate the way in which electoral manifestos might refer to policies. For example, a literal translation of the term 'minimum wage' from English (source language) to French (target language) would miss the fact that in France the minimum wage is frequently referred to with its acronym (*SMIC*).[4] Similarly, in Germany, discussion of abortion rights in electoral manifestos often make references to the respective articles in the penal code (§ 218- §219). While an internet and news media search can allow researchers to tease out policy-specific signifiers, it is important to validate the keyword selection with native speakers, whose country-specific and substantive knowledge of the way in which policies are discussed in the political discourse of different countries can increase the credibility of the results. For this context, it is therefore not sufficient to speak a language, but it is also important to be be acquainted with the political context, in order to have insights on the way in which parties might discuss these policies in their electoral manifestos. Specifically for this analysis, three native speakers were asked to validate the keyword selection (one per language).

[4] SMIC: Salaire minimum interprofessionnel de croissance

Validation checks between coders and a multilingual dictionary analysis

From a linguistic perspective, it is important to note that the way in which keywords are represented syntactically in the text might vary across languages as well as topics. While translating and retrieving information on word-level n-grams, such as ’minimum wage’, is a straightforward process, keywords might also appear within the same sentence, but may not appear contiguous to each another. Consider the following excerpts from two Irish electoral manifestos on the question of same-sex marriage: 1. ’Labour is committed to holding a referendum to provide for constitutional recognition of same-sex marriage.’[5] and: 2. ’[...] allow same-sex couples [to] enjoy the rights and responsibilities of civil marriage’[6]. These are both sentences, which clearly refer to the topic of same-sex marriage legislation, but while a keyword specification designed to retrieve contiguous keywords only (’same-sex marriage’) would correctly identify the first text excerpt, it would fail to identify the second excerpt. Where it is necessary to do so, it is therefore important to specify within the search criteria that keywords need not be contiguous. Familiarity with regular expressions (RegEx) as well as language-specific validation checks constitute particularly important steps for this. This holds especially true for synthetic languages, such as German, where word form changes tend to occur through the use of affixes or other internal modifications within the words themselves. Once again, the importance of validation in this context cannot be overstated.

The most time-consuming aspect of multilingual dictionary construction is that of keyword evaluation. When evaluating the policy-specific multilingual dictionary, it is important to answer two questions: First, do the keywords present in the policy-specific multilingual dictionaries correctly identify the text excerpts in which the respective policies are actually discussed (*classifier precision)*? Second, to what an extent do the multilingual dictionaries fail to identify instances in which the respective policies are mentioned in the text (*classifier recall)*? In order to validate the policy-specific multilingual dictionary, I have extracted from the Manifesto Project portal all quasi-sentences which were assigned the CMP codes 412, 603, 604, for all parties competing in Belgium (Francophone electoral manifestos only), France, Germany, Ireland and the United Kingdom from the year 2000 onward. This was done via the R wrapper package (*manifestoR*) for the

[5] Ireland, Labour Party electoral manifesto (2011), *One Ireland - Jobs, Reform, Fairness*

[6] Ireland, Green Party electoral manifesto (2007), *Manifesto 2007 - It’s Time*

CMP's API (Lewandowski et al., 2020). The decision to use quasi-sentences assigned to the policy areas listed above is borne of the fact that, based on the CMP coding scheme, discussions of the policies which this article seek to uncover are more likely to be present in quasi-sentences coded as belonging to these CMP policy codes (for further details see Table A1 in the Appendix).

To validate this corpus (which contains 4,765 quasi-sentences), three native speakers per language were asked to read all quasi-sentences, which were coded with the CMP codes described above, and note whether the quasi-sentences referred to either the minimum wage, to abortion or same-sex marriage or were unrelated to any of these topics. To ensure that coders were not unduly influenced, the text excerpts that the coders received did not contain any information on how the custom-built policy-specific multilingual dictionary or indeed how the other coders had classified the text. These validations were then compared to one another. Agreements between coders in classifying quasi-sentences was quite high (Krippendorff's *alpha*: 0.903). Subsequently, I compared the human interpretation of the text with the results yielded by the multilingual dictionary which had been constructed previously. Human validation of automated computer-assisted output constitutes, as is well known, a fundamental step in quantitative text analysis. A human interpretation of the raw text thus remains the 'gold standard', the benchmark, against which to compare the results derived from machine-run text analyses (Grimmer & Stewart, 2013; Rauh, 2018).

In order to assess the performance of the multilingual dictionary, I aggregate - based on majority rule - the coding assigned by the three coders to the quasi-sentences of electoral manifestos into one score. This means that every quasi-sentence in the validation corpus is associated with a binary score, which measures whether, according to the coders, the text excerpt mentions a certain policy ($N = 427$) or not ($N = 4,338$). The results presented in Table 1 illustrate the performance of the multilingual dictionary on the subset of the CMP corpus, which was used for validation purposes. In the table below, *precision* measures the number of - as assessed by human coders - true quasi-sentences featuring the policies of interest as a share of the number of quasi-sentences predicted to include these policies. *Recall* measures the number of quasi-sentences which are classified by the multilingual policy-specific dictionary as containing the policies of interest against the true number of quasi-sentences which are classified as containing these policies. This measure thus accounts for 'missed' quasi-sentences

Country	Measure	Minimum wage *(CMP code: 412)*	Abortion *(CMP code: 603-604)*	Same-sex marriage *(CMP code: 603-604)*
France & Belgium	*Precision*	0.88	0.95	0.75
	Recall	1.00	0.93	1.00
	F-measure	0.94	0.94	0.86
Germany	*Precision*	0.94	0.95	0.81
	Recall	0.98	0.71	0.77
	F-measure	0.96	0.81	0.79
Ireland & UK	*Precision*	0.97	0.86	0.77
	Recall	0.95	0.81	0.77
	F-measure	0.96	0.84	0.77

Table 1: Validation of the multilingual dictionary on the selected corpus of the CMP.

(or false negatives). Finally, the F_1 score represents the harmonic mean of precision and recall and is measured as:

$$F_1 = \frac{2 * (precision * recall)}{precision + recall} \tag{1}$$

As shown by the high values registered in these metrics, it is clear that the multilingual policy domain-specific dictionary is able to retrieve most relevant text excerpts. While Table 1 illustrates that across all policies the performance of the multilingual dictionary is generally quite satisfactory, it is noteworthy that the results are particularly good for text excerpts classified as discussing minimum wages. This is because across the different languages which are being analyzed here, the term tends to appear as single or two contiguous keywords. Unsurprisingly, the more precisely a policy can be expressed, the higher is the precision value, as recorded in the validation of the multilingual dictionary. For a term such as abortion there are a variety of different terms, some of which can be coloured by distinctly political considerations (e.g. pro-life, ’unborn children’), which can be used to express this concept. This, of course, makes the risk of omitting relevant sentences higher. Nevertheless, as illustrated by the performance metrics shown in Table 1 these concerns seem to be largely allayed in this analysis.

Policies instead of policy areas: Applying the multilingual dictionary analysis to the corpus of the CMP

The positive results shown in the performance metrics illustrated above provide encouraging indications that multilingual dictionary analysis applied to the CMP corpus can yield satisfactory results. But what is the added value of utilizing policy-specific multilingual dictionaries for researchers interested in comparative electoral politics? Based on the coding scheme provided by the CMP, quasi-sentences referring to minimum wages, abortion and same-sex marriage could, as discussed, be assigned into a series of different CMP-defined policy areas. Are policies, however, really classified where one would expect them to be? By searching for the same keywords that have yielded satisfactory results in the validation exercise described above, I find that quasi-sentences which explicitly mention the policies of interest can frequently appear categorized as belonging to different CMP-defined policy areas.

Figure 2 illustrates the extent to which the policies examined in this article are classified into different CMP policy areas. This is based on an analysis of the electoral manifestos of all parties competing in general elections in Austria, Germany, Ireland, Switzerland and the UK from 2000 onward. For ease of interpretation CMP codes have been collapsed into the seven macrodomains, in which CMP codes are divided in. As is it is clear, the same policy can be assigned to very different CMP policy areas. For example, while most quasi-sentences which deal with minimum wages are assigned to the CMP code 412, which includes those quasi-sentences, which show 'support for direct government control of economy (e.g. control over prices, introduction of minimum wages)', one can also note that, based on the multilingual policy-specific dictionary, approximately 50% of quasi-sentences which discuss minimum wages are coded in other policy areas.

Moreover, as noted in the introduction, CMP codes encompass a number of different policies and topics. The validation exercise presented above finds that only 8.9% of quasi-sentences actually discuss the policies of interest. Again, it is worth noting that these results are conditioned by the fact that the guiding principle of the CMP's goal-oriented coding scheme is that of coding quasi-sentences as belonging to certain policy areas if statements are being made about specific policy objectives, as opposed to policies. As an example,

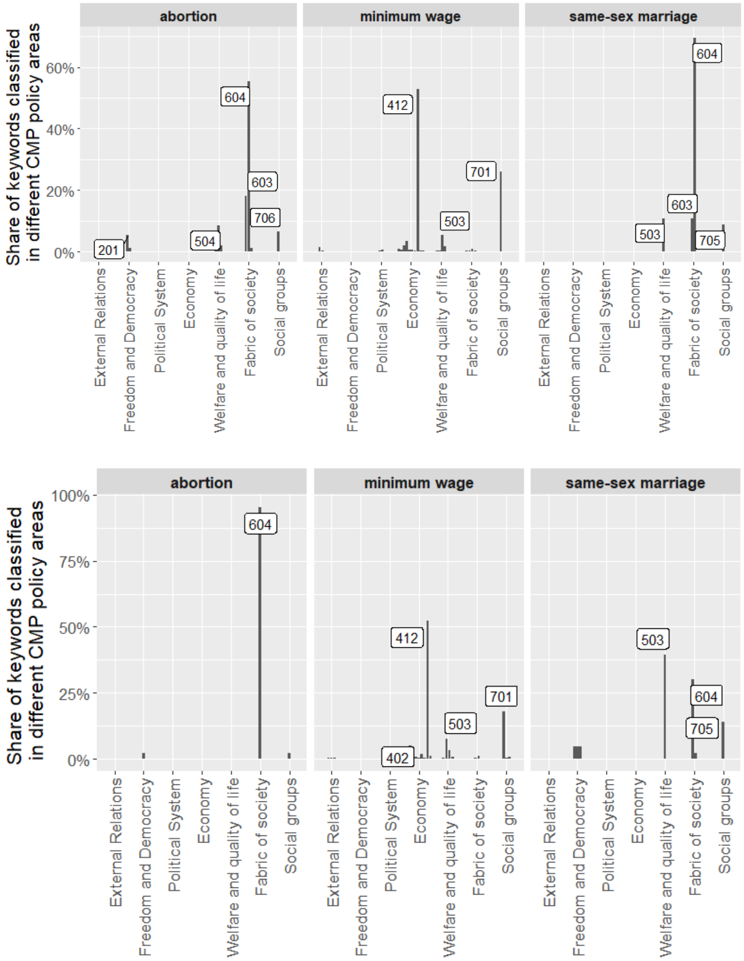

Figure 2: The assignment of keywords in different CMP policy areas, collapsed into different domains, as recorded in Austrian, German and Swiss (above) and British and Irish (below) electoral manifestos (2000-2021)

the following quasi-sentence, which is taken from the Scottish National Party's 2015 electoral manifesto states that 'by increasing the minimum wage and supporting a fair wage economy we can increase the disposable income of low and middle-income households'[7] is coded as belonging to the CMP policy area 'Equality' (*per503*), which includes all quasi-sentences which

[7] United Kingdom, Scottish National Party (SNP) electoral manifesto (2015), *Strong for Scotland*

make claims 'about the need for the fair treatment of all people'. In this excerpt, the minimum wage is thus seen as a policy tool, as a mean, to establish a fairer economy.

An application: Electoral salience and keywords-in-context

The analysis presented thus far has illustrated the discrepancies that can be found when comparing the output derived from a custom-built policy-specific multilingual dictionary to the output yielded from an analysis, which only utilizes CMP codes. In this section, I focus on two concrete applications of multilingual policy-specific dictionaries, which might be of interested for scholars of electoral politics: 1) electoral salience and 2) the broader context in which policies are discussed in electoral manifestos. As noted above, researchers using CMP data have frequently used the number of quasi-sentences assigned to a policy area as a proxy for the salience that a certain policy area plays in electoral manifestos. A similar analysis can be applied here by focusing on the number of natural sentences which are devoted to a given policy as a *share* of all natural sentences contained in an electoral manifesto: the higher the number of sentences discussing a certain policy, the higher its electoral salience. This policy-specific analysis has two advantages. First, by relying on the prior annotation of CMP electoral manifestos, researchers are constrained to using the quasi-sentence as a unit of analysis; this, however, might not be ideal as quasi-sentences focus on capturing the objective of a text excerpt, as opposed to the actual policy which is being discussed. Second, it is important to note that only a selection of the electoral manifestos contained in the CMP database are annotated and subdivided into quasi-sentences. An analysis of electoral data, which would use quasi-sentences as its units of analysis might thus omit several potentially relevant electoral manifestos from its analysis. This is why in this analysis I use natural sentences as opposed to quasi-sentences.

Following from the text analysis presented above, I obtain a multilingual subset of the CMP corpus which contains all natural sentences mentioning the policies, which I examine in this article. In Figure 3, I illustrate the salience of these policies by plotting the share of natural sentences discussing the different policies as a share of all natural sentences contained in a manifesto against parties' partisanship, as measured by the Chapel Hill Expert Survey (CHES) left-right scale. Each dot in Figure 3 represents the electoral salience that these policies play for a particular party in a particular year. While, one can see that left-wing parties are more likely to discuss abortion, same-sex marriage and the minimum wage in their electoral manifestos, I also find support for the fact that radical right parties are likely to apportion a greater share of their electoral manifesto to discussions of abortion and the minimum wage than centre and centre-right parties. While a country-specific analysis of these parties' discourse on these issues is not within this article's scope, it is interesting to note that at least for two of these policies issue salience seems to follow a curvilinear relationship.

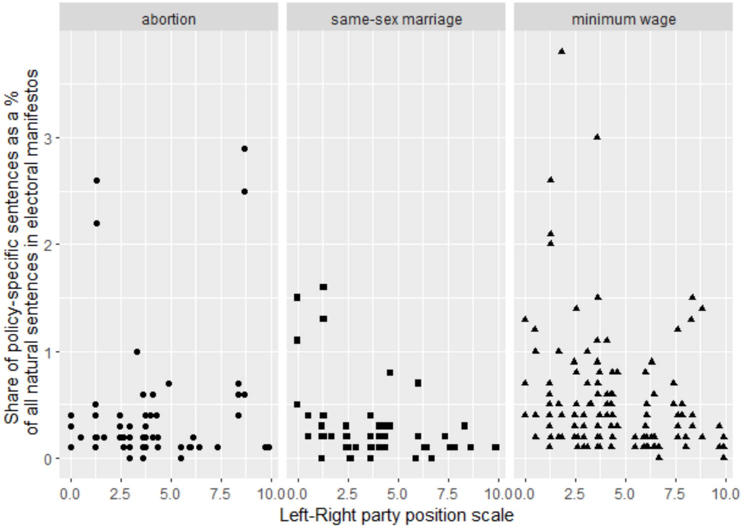

Figure 3: Share of policy-specific natural sentences as a share of all natural sentences in the electoral manifestos of Austria, Belgium, France, Germany, Ireland, Switzerland, United Kingdom (2000-2021). Each dot represents electoral salience for a given party in a given year.

While it is certainly interesting and important to consider the salience which certain policies play in electoral manifestos, analysts might also profit from understanding the context in which parties discuss policies. I therefore proceed in illustrating the context in which political parties discuss the policies considered here. In order to do this I first translate the corpus of all sentences which discuss the policies of interest in the electoral manifestos into a common 'pivot' language, that is English. I do this so that readers who do not read French and German (or any other language which analysts might be interested in exploring) can also understand the context in which policies are discussed. Given the size of the text as well as the obvious time and financial advantages provided by a software translation over a human translation, I translated the corpus into English with Google Translate. The usage of Google Translate in quantitative text analyses applied to political texts has been found to be a valid and useful approach for bag-of-words models, which also results in high levels of agreements between 'gold standard' human translations and machine translations (De Vries et al., 2018). Moreover, from a practical perspective, recently, it has become possible to upload the documents one wishes to translate directly onto Google Translate. As the website supports different file formats, researchers can transform their multilingual corpora extracted from the CMP corpus into .xlsx files and then upload these directly to the website. In line with best practices, before translating the text, I have made sure to pre-process the data, by removing tags and symbols which might negatively influence the quality of the translation (Dashtipour et al., 2016). A sample of the translated output has then been examined by two native speakers and has been judged satisfactory.

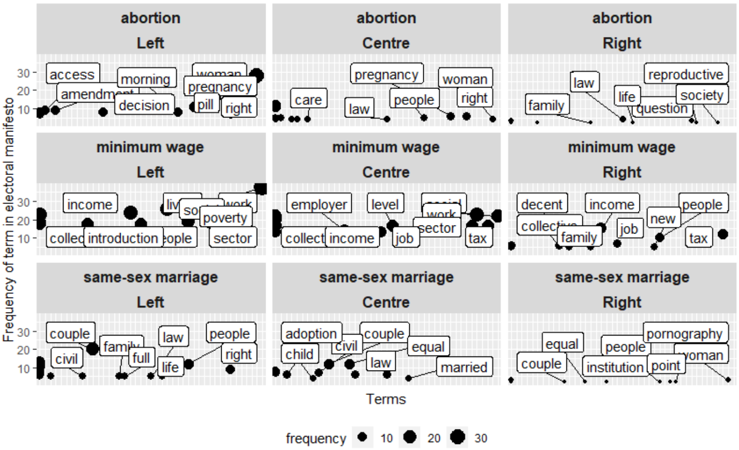

Figure 4: Keywords-in-context analysis: Most commonly used nouns and adjectives used for the discussion of policies by parties competing in Austria, Belgium, France, Germany, Ireland, Switzerland, United Kingdom (2000-2021)

For the purposes of this article, indications of the immediate context in which the policies are discussed is given by the most frequent words which appear in the same sentence, in conjunction with discussions of the policies analyzed here. For this application, I focus on examining the most frequently used nouns and adjectives present in the sentences in which the policies examined here appear in the text of electoral manifestos. I identify the most frequently employed nouns and adjectives by implementing the part-of-speech (POS) tagging algorithms contained in R's *udpipe* package

(Wijffels, 2022). For the visualization of the results, I code partisanship by transforming the continuous left-right scale into a tripartite categorization constituted by left, centre, and right-wing parties. The analysis displayed in Figure 4 shows that left-wing parties frequently mention words such as ’poverty’, ’social’ and ’living’ in the context of discussions on the minimum wage; thus illustrating the social dimension of discussions on this policy. Centre and right-wing parties instead tend to also focus on words such as ’tax’ and ’employer’. In the context of access to abortion, left-wing parties prioritize words like ’access’, ’right’, ’morning-after pill’ while right-wing parties frequently use terms such as ’life’ and ’family’. Again, scholars who would be interested in examining the substantive differences which exist between specific party families or countries would be well advised to subset this analysis on a country or party-level. As this article seeks to make a methodological contribution, the categorization has been kept purposefully broad. The application presented here has thus shown that the multilingual dictionary analysis employed in this article can aid in understanding the electoral salience of these policies as well as the broader context in which these policies are discussed.

Conclusion

This study was motivated by the many articles published by researchers interested in comparative electoral politics who make use of the CMP. While the usage of CMP policy area codes is practical and often helpful, the categorization offered by the CMP reaches its limits when researchers are interested in *policies* as opposed to CMP-defined policy areas. The electoral manifestos included in the CMP are coded with the aim of capturing the policy objectives present in a text excerpt and do, therefore, not necessarily concentrate on policies in and of themselves. Here, I have argued that through the use of a multilingual dictionary analysis, researchers can exercise greater control by focusing on the policies that they are interested in as opposed to be being constrained to a selection of quasi-sentences identified by the CMP’s coding scheme. By considering how three electorally salient (and divisive) policies have been discussed across the party-political spectrum of seven different countries in three different languages, this article has illustrated the extent to which multilingual policy domain-specific dictionaries, with few keywords, are able to retrieve most instances in which policies are discussed. Validation checks are intuitive and compared to other validation endeavours (e.g. topic or sentiment dictionaries) less time-consuming. Here, I have shown that the keywords linked to policies are specific enough to retrieve and recall a satisfactory number of text excerpts. Compared to topic dictionaries, it thus seems reasonable to conclude that there is a lower risk that this approach retrieves false positives, misclassifies keywords or fails to correctly identify text excerpts which discuss relevant policies.

In this article, I have first illustrated empirically that specific policies might appear categorized in a series of different CMP policy areas/codes. This, I have argued, should motivate comparative researchers who are interested in understanding the way in which different political parties discuss policies in their electoral manifestos to use a relatively simple technique from the quantitative text analysis toolbox, that is multilingual dictionaries. In the second part of this article, I have used the resulting corpus to explore the electoral salience of policies as well as the context in which policies are discussed. I have shown that by only focusing on specific policies the size of the corpus can be significantly reduced and the translation of the multilingual text into a common language becomes easy, economical (if not free) and computationally simple. Furthermore, by leveraging POS tagging, it becomes straightforward to understand the most frequent words which are used in conjunction with the policies analysts are interested in examining. This can provide substantive indications of the way in which political parties frame policies in their electoral manifestos.

References

Akkerman, T. (2015). Gender and the radical right in western europe: A comparative analysis of policy agendas. *Patterns of Prejudice*, *49*(1-2), 37–60.

Albaugh, Q., Sevenans, J., Soroka, S., & Loewen, P. J. (2013). The automated coding of policy agendas: A dictionary-based approach. *6th Annual Comparative Agendas Conference, Atnwerp, Beligum.*

Allen, N., & Bara, J. (2019). Marching to the left? programmatic competition and the 2017 party manifestos. *The Political Quarterly*, *90*(1), 124–133.

Baden, C., & Stalpouskaya, K. (2015). Common methodological framework: Content analysis. a mixed-methods strategy for comparatively, diachronically analyzing conflict discourse. *Ludwig Maximilian University Munich: INFOCORE Working Paper*, *10*, 2015.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital journalism*, *4*(1), 8–23.

Budge, I., & Farlie, D. (1983). *Explaining and predicting elections: Issue effects and party strategies in twenty-three democracies*. Taylor & Francis.

Chueri, J. (2022). An emerging populist welfare paradigm? how populist radical right-wing parties are reshaping the welfare state. *Scandinavian Political Studies*, *45*(4), 383–409.

Cova, J. (2023). Politicizing the minimum wage: A multilingual text analysis of minimum wages in european electoral manifestos. *Journal of European Social Policy*, 09589287231199561.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive computation*, *8*(4), 757–771.

Däubler, T., Benoit, K., Mikhaylov, S., & Laver, M. (2012). Natural sentences as valid units for coded political texts. *British Journal of Political Science*, *42*(4), 937–951.

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430.

Dietze, G., & Roth, J. (2020). *Right-wing populism and gender: European perspectives and beyond*. transcript Verlag.

Engeli, I., Green-Pedersen, C., & Larsen, L. T. (2012). *Morality politics in western europe: Parties, agendas and policy choices*. Springer.

Farstad, F. M. (2018). What explains variation in parties' climate change salience? *Party Politics*, *24*(6), 698–707.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Hooghe, L., Marks, G., & Wilson, C. J. (2002). Does left/right structure party positions on european integration? *Comparative political studies*, *35*(8), 965–989.

Horn, A., Kevins, A., Jensen, C., & Kersbergen, K. V. (2017). Peeping at the corpus–what is really going on behind the equality and welfare items of the manifesto project? *Journal of European Social Policy*, *27*(5), 403–416.

Klingemann, H.-D., Hofferbert, R., Budge, I., Keman, H., Keman, I., Bergman, T., Pétry, F., Strom, K., et al. (1994). *Parties, policies, and democracy*. Westview Press.

Lawlor, A. (2015). Local and national accounts of immigration framing in a cross-national perspective. *Journal of Ethnic and Migration Studies*, *41*(6), 918–941.

Lawlor, A., & Tolley, E. (2017). Deciding who's legitimate: News media framing of immigrants and refugees. *International Journal of Communication*, *11*, 25.

Lewandowski, J., Merz, N., & Regel, S. (2020). *Manifestor: Access and process data and documents of the manifesto project* [R package version 1.5.0]. https://CRAN.R-project.org/package=manifestoR

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). Computational communication science| when the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, *13*, 21.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277.

Mikhaylov, S., Laver, M., & Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, *20*(1), 78–91.

Pearson, K., & Dancey, L. (2011). Speaking for the underrepresented in the house of representatives: Voicing women's interests in a partisan era. *Politics & Gender*, *7*(4), 493–519.

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, 825–850.

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, *44*(1), 97–131.

Protsyk, O., & Garaz, S. (2013). Politicization of ethnicity in party manifestos. *Party Politics*, *19*(2), 296–318.

Rauh, C. (2018). Validating a sentiment dictionary for german political language—a workbench note. *Journal of Information Technology & Politics*, *15*(4), 319–343.

Rubin, O., Baekkeskov, E., & Öberg, P. (2021). A media visibility analysis of public leadership in scandinavian responses to pandemics. *Policy Design and Practice*, *4*(4), 534–549.

Schulten, T. (2014). *Minimum wage regimes in europe*. Berlin: Friedrich-Ebert-Stiftung. Available at http://library. fes. de/pdf . . .

Volkens, A., Burst, T., Krause, W., Lehmann, P., Regel, S., & Zehnter, L. (2021). The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2021a. https://doi.org/10.25522/manifesto.mpds.2021a

Vos, D., & Van Aelst, P. (2018). Does the political system determine media visibility of politicians? a comparative analysis of political functions in the news in sixteen countries. *Political Communication*, *35*(3), 371–392.

Wijffels, J. (2022). *Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'udpipe' 'nlp' toolkit* [R package version 0.8.9]. https://CRAN.R-project.org/package=udpipe

Zulianello, M. (2014). Analyzing party competition through the comparative manifesto data: Some theoretical and methodological considerations. *Quality & Quantity*, *48*(3), 1723–1737.

Minimum wage	**Abortion**	**Same-sex marriage**
CMP Code (412): Support for direct government control of economy. • Control over prices • Introduction of minimum wages	*CMP Code (604)*: Traditional Morality (negative). • Support for divorce, abortion etc • General support for modern family composition • Calls for the separation of church and state. *CMP Code (603)*: Traditional Morality (positive): • Prohibition, censorship and suppression of immorality and unseemly behaviour • Maintenance and stability of the traditional family as a value • Support for the role of religious institutions in state and society.	*CMP Code (604)*: Traditional Morality (negative). • Support for divorce, abortion etc • General support for modern family composition • Calls for the separation of church and state. *CMP Code (603)*: Traditional Morality (positive): • Prohibition, censorship and suppression of immorality and unseemly behaviour • Maintenance and stability of the traditional family as a value • Support for the role of religious institutions in state and society. *CMP Code (503)*: Equality (positive) • Special protection for underprivileged social groups • Removal of class barriers • The end of discrimination (e.g. racial or sexual discrimination)

Table A1: The way in which different policies are categorized in the CMP's coding scheme (V5)

Language	Keyword
German	mindestlo*
English	minimum wag*
	living wag*
French	smic
	salair* minim*

Table A2: Keyword specification: Minimum wage

Language	Keyword
German	[gleichgeschlecht* | lgbt* | schwul* |homosex* | gay] & ehe
	[gleichgeschlecht* | lgbt* | schwul* |homosex* | gay] & heirat*
	ehe für alle | homo-ehe
English	[same-sex* | same sex* | homosex* |equal* |lgbt* | gay] & marr*
	marriage for all
French	[homosex* | lgbt* | meme sex* |egal* | gay] & mari*
	mariage pour tous

Table A3: Keyword specification: Same-sex marriage

Language	Keyword
German	[abbr* | entsch* | besti*] & schwangersch*
	schwangerschaftsabb*
	abtreib*
	pro-life
	fristenlös*
	fristenregel*
	pro-choice
	pill* danach
	§218, §219
English	[terminat* | decid* | decis*] & pregnan*
	abort*
	reproductiv* & right*
	pro-life
	pro-choice
	morning after pill
	[eighth | 8th] & amendment
French	avort*
	ivg
	loi veil
	[pro-life | pro-vie]
	pro-choi*
	pilule du lendem*
	interrup* & grosses*

Table A4: Keyword specification: Abortion