

# Speaking Their Language?: Multilingualism in Party Communication across Democracies\*

Taishi Muraoka<sup>†</sup> Dahjin Kim<sup>‡</sup>

Christopher Lucas<sup>§</sup> Jacob Montgomery<sup>¶</sup> Margit Tavits<sup>||</sup>

**Short title:** Multilingualism in Party Communication

**Keywords:** Language, Multilingualism, Political Party, Social Media

---

**\*Acknowledgments:** We thank Chapman Rackaway, Editor Graeme Robertson, and three anonymous reviewers for their helpful comments. We are also grateful to Afiq bin Oslan and Amid Singh and the many research assistants around the world hired through Upwork. Previous versions of this work were presented to helpful audiences at the Department of Quantitative Theory and Methods at Emory University, the Political Methods Workshop at Vanderbilt University, and the Department of Political Science at Ohio State. Funding for this project was provided by the Weidenbaum Center on the Economy, Government, and Public Policy, the National Science Foundation (#2215008), and the Carnegie Corporation of New York (G-23-60440).

<sup>†</sup>Address: Institute of Political Science, Academia Sinica, 128 Academia Rd, Sec. 2, Nangang, Taipei 115, Taiwan. Email: tmuraoka@gate.sinica.edu.tw

<sup>‡</sup>Address: Department of Political Science, Washington University in St. Louis., One Brookings Dr, St. Louis, MO 63130, United States. Email: dahjin.kim@wustl.edu

<sup>§</sup>Address: Department of Political Science, Washington University in St. Louis., One Brookings Dr, St. Louis, MO 63130, United States. Email: christopher.lucas@wustl.edu

<sup>¶</sup>Address: Department of Political Science, Washington University in St. Louis., One Brookings Dr, St. Louis, MO 63130, United States. Email: jacob.montgomery@wustl.edu

<sup>||</sup>Address: Department of Political Science, Washington University in St. Louis., One Brookings Dr, St. Louis, MO 63130, United States. Email: tavits@wustl.edu

**Article Title:** Speaking Their Language?: Multilingualism in Party Communication across Democracies

**Abstract:** Which parties embrace multilingualism in their communication? Despite growing interest in parties' multilingualism among normative scholars of deliberative democracy, empirical research has largely overlooked the linguistic aspect of party competition. We leverage large-scale data on Facebook posts by more than 800 parties in 87 democracies and analyze their day-to-day language practices. By so doing, we develop, for the first time, the classification of monolingual and multilingual parties around the world. Moreover, using this novel dataset, we explore what factors are associated with parties' adoption of multilingualism and how multilingual parties predict the language use of candidates they nominate. Overall, this study provides the most comprehensive picture of parties' multilingualism in contemporary democracies and sets agendas for future research in the intersection of parties and language representation.

**Replication Materials:** The data, code, and any additional materials required to replicate all analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: <http://dx.doi.org/10.7910/DVN/K3WTHS>.

**Word Count:** 9368

Linguistic diversity poses unique challenges to democracy. Major language differences among citizens make it more difficult for them to enter into meaningful political dialogues and develop a common public sphere. The failure to create a shared deliberative forum results in limited opportunities for different language groups to exchange their viewpoints and engage in rational persuasion. Eventually, this circumscribed deliberative process can have downstream consequences on economic growth, public goods provision, and even civil conflict (Desmet, Ortuño-Ortín and Wacziarg 2012; Liu and Pizzi 2018).

Responding to the deliberative challenges in linguistically diverse democracies, several theorists claim that multilingual parties can be important intermediaries of deliberation among different language groups (Bonotti and Stojanović 2022; Stojanović and Bonotti 2020).<sup>1</sup> By operating in multiple languages, these parties create a public sphere in which citizens from different language groups are informed about each other's perspectives, engage in constructive debates, and pursue political projects that achieve shared goals. By so doing, multilingual parties can uphold more inclusive and respectful democratic practices than monolingual parties that represent a single language group and reinforce social cleavages.

Despite the normative importance of multilingual parties, there is very little empirical understanding of parties' multilingualism in democracies around the world. Indeed, prior studies have looked into only one or a few countries to understand the language practices of political parties (Caluwaerts and Reuchamps 2014; De Bres, Rivera Cosme and Remesch 2020; Rubin 2014).<sup>2</sup> Thus, Bonotti and Stojanović (2022) deplore "[i]n our

---

<sup>1</sup> In the SI Section A (pp. 1-2), we discuss how parties' multilingualism is different from other related concepts, such as multiculturalism (Westlake 2018), cross-ethnic appeals (Devasher and Gadjanova 2021; Gadjanova 2021), and ethnic mobilization (Chandra 2007; Strijbis and Kotnarowski 2015).

<sup>2</sup> For studies on politicians' language use, see Crisp, Demirkaya, Schwindt-Bayer and

survey of the literature on multilingualism and political parties, [...] we were also struck by the relative lack of empirical research on the linguistic dimensions of party life” (p. 480). As a result, we do not have answers to even the most fundamental questions about parties’ multilingualism. Which parties use multilingual appeals? What explains their adoption of multilingualism? And how do multilingual parties relate to the campaign communications of candidates they recruit? Answering these questions has important implications for how deliberative processes work in different democracies.

We argue that the limited attention to the linguistic dimension of party competition stems from the lack of appropriate methodological tools. Thus, the primary goal of this article is to resolve this issue by analyzing a novel dataset of parties’ communications on Facebook, a widely-used platform for parties in much of the world. As social media posts represent parties’ direct attempts to communicate with citizens on a daily basis, analyzing these posts enables us to have a natural assessment of how parties choose to communicate in different languages. Methodologically, we draw on recent advances in computational models for language detection to generate a dataset on parties’ multilingualism in 87 countries. Crucially, our data encompasses not only well-studied multilingual countries, like Canada and Switzerland, but also less-studied ones, such as Malaysia and Lesotho.

After establishing the monolingual/multilingual classification of more than 800 parties, we analyze what factors are associated with their adoption of multilingualism. In answering this question, we situate our argument within the incentives-constraints model of seat-maximizing parties (Abou-Chadi, Green-Pedersen and Mortensen 2020; Tavits and Potter 2015; Toubeau and Wagner 2016). This theoretical framework suggests that parties should have greater incentives to use multilingual appeals as linguistic diversity increases. However, whether this strategy is feasible is conditioned by two types of constraints: (1) institutional constraints, which shape parties’ calculus regarding whether they need to appeal across language boundaries and build a broad coalition to

---

Millian (2018) and Ringe (2022).

gain a seat (Calvo and Hellwig 2011; Cox 1990; Horowitz 1985); and (2) ideological constraints, which shape parties' expectations about whether multilingual appeals would reinforce or undermine their ideological brands (Adams and Somer-Topcu 2009; Lupu 2014; Tavits 2007).

Our theory of institutional constraints suggests that the association of linguistic diversity with parties' adoption of multilingualism is greater under majoritarian electoral systems than under proportional ones. This occurs because majoritarian systems necessitate parties to build a broad winning coalition to gain a seat in the district, while proportional systems assure a more proportional translation of votes into seats. This means that as linguistic diversity increases, the extent to which minority language groups become decisive in determining who gains a seat increases at a higher rate under majoritarian systems than proportional systems. As a result, when linguistic diversity is high, the former requires parties to cross language lines and appeal to different language groups more than the latter. Our empirical results generally support these expectations, showing modest evidence that increases in linguistic diversity are more strongly associated with parties' adoption of multilingualism under majoritarian systems than under proportional systems.

Our argument about the ideological constraints, in turn, suggests that the association between linguistic diversity and parties' adoption of multilingualism should be greater among socially left-leaning parties. This is because these parties do not face any branding problems by accepting multilingualism, as committing to minority inclusion is consistent with their existing ideological brands. By contrast, it is more difficult for socially rightist parties to use multilingual appeals because their supporters may perceive such appeals as a betrayal of the party brand. In line with these expectations, we demonstrate that, as linguistic diversity increases, socially left-leaning parties are more likely to adopt multilingual appeals than their right-leaning counterparts. Moreover, a simple placebo test shows that this pattern does not hold as clearly when focusing on parties' economic left-right ideology.

We extend our analysis to individual candidates and examine how multilingual parties relate to candidate-level language practices. To do so, we analyze the Facebook posts of candidates for lower house elections in twelve multilingual countries that took place between 2020 and 2023. We show that candidates of multilingual parties are more likely to embrace multilingualism and use a minority language than candidates nominated by monolingual parties. These findings indicate that parties' multilingualism reflects not their symbolic gestures to appeal to different language groups but rather their deeper commitment to enhancing language-based representation.

In total, this study offers the most comprehensive picture of parties' multilingualism in contemporary democracies – how it manifests within party systems and the institutional and ideological factors shaping its adoption. In addition, we advance and validate a novel approach to identifying language use in large collections of political texts. By so doing, we provide a valuable new measure of parties' multilingualism in democracies across the globe, which opens up new avenues of future research in the intersection of language and politics.

### **Language, deliberation, and multilingual parties**

Like race, ethnicity, and gender, language is a core component of social identity that defines political cleavages (Barrington 2022; Laitin 1989; Marquardt 2022; Medeiros 2017). Treating language as a part of identity, many empirical studies have examined how language shapes political preferences and behavior (Frye 2015; Lee and Pérez 2014; Ricks 2020; Pérez and Tavits 2019; Zárata, Quezada-Llianes and Armenta 2024).<sup>3</sup> In some countries, such as Belgium and Spain, linguistic divides overlap with other social

---

<sup>3</sup> Pérez and Tavits (2022) show that language is important to understand other, non-identity based political attitudes, including gender equality, environmental protection, and policy priorities.

divisions (e.g., ethnicity). In these cases, language is seen as a primary factor that reinforces cultural differences and political cleavages between groups (Kulyk 2011; Laitin 1989). In other countries, such as India, Lesotho, and Malaysia, the use of a certain language (e.g., English) offers socioeconomic and other advantages. In these cases, language introduces cross-cutting social divisions and identities, generating a mosaic of groupings that together map onto political loyalties.

However, the political importance of language goes beyond its role as an identity marker. In recent decades, normative theorists have focused on the instrumental role of language to channel political deliberation (Lacey 2017; Schmidt 2014; Strani 2020; Young 1990). Their central debate has been whether linguistic diversity hinders effective deliberation in the public sphere. For some scholars, the presence of multiple languages in a political system constitutes an obstacle to the formation of a common public sphere (Addis 2007; Lacey 2017). After all, language differences encourage different language groups to engage in political debates within their group boundaries, which reduces the chance that different parts of the system enter into a meaningful dialogue. According to this view, having a common language is a prerequisite for functioning deliberative democracy.<sup>4</sup>

By contrast, other theorists advocate multilingualism in democratic deliberation. For example, Strani (2020) defends multilingualism on the grounds that monolingual public spheres are exclusionary. Allowing only one language in public debates means creating language hierarchies, and this inevitably results in marginalizing minority languages. Similarly, Schmidt (2014) supports multilingual practices in public deliberation because they can promote more inclusive and egalitarian participation among all citizens.<sup>5</sup> For

---

<sup>4</sup> Consistent with this claim, some studies find that a *lingua franca* has a positive effect on intergroup relationships and economic development (Kumove 2022; Liu 2015; Liu and Pizzi 2018).

<sup>5</sup> Specifically, Schmidt (2014) points out three advantages of multilingual public spheres: (1) the engagement of different language communities results in the most legitimate form

him, forcing citizens to use a common language is to require some of them to change who they are. Allowing multilingual interactions in public spheres should achieve more legitimate deliberation outcomes.

In modern democracies, political parties play critical roles in structuring public deliberation, and multilingual parties become particularly vital to creating multilingual public spheres (Bonotti and Stojanović 2022; Stojanović and Bonotti 2020).<sup>6</sup> For one, as primary vehicles of representation, parties shape which language enters the process of political deliberation. Simultaneously engaging with different language groups, multilingual parties ensure that political debates take place beyond the boundaries set by the linguistic border. Second, parties also become important informational and educational sources for voters. By providing the same information in different languages, multilingual parties enhance equal access to information, provide opportunities to learn about different viewpoints, and enable constructive dialogues across language groups.

The importance of multilingual parties in sustaining meaningful democratic deliberation in linguistically diverse settings is nicely illustrated by the Belgian party system. Belgium's institutional configurations enable parties to win votes by appealing solely to one language group. As Caluwaerts and Reuchamps (2014) describe, this reduces their incentives to communicate to citizens in the other language group and erodes lines of communication across language lines. The separation of the party system by language eventually distorts the deliberative capacities of the entire system, leading to political crises characterized by tense communal relations and political instability. In this way, the Belgian case illustrates how the presence of mostly monolingual parties,

---

of governance (legitimation advantage); (2) the inclusion of all language groups in the deliberation table leads to truly "common" decisions (common good advantage); and (3) the deliberation among different language groups enhances human capacity by requiring us to see the world from others' points of view (human flourishing advantage).

<sup>6</sup> Media is equally important to shape public spheres (Caluwaerts and Reuchamps 2014).

each representing a single language group, may not be sufficient to construct a strong public sphere in multilingual democracies.

The above discussions suggest that there are extensive normative debates about the relationship between linguistic diversity and deliberation and how multilingual parties play a part in this process. However, empirical investigation of which parties actually adopt multilingualism is limited to a handful of case studies that have analyzed parties' language use based on anecdotes or qualitative reading of parties' election programs (Caluwaerts and Reuchamps 2014; De Bres et al. 2020; Rubin 2014; Stojanović and Bonotti 2020). This gap between normative theories and empirical research is striking and requires a more systematic assessment of parties' multilingual practices. To do so, we develop a method of detecting multilingual parties by applying computational methods of language detection to parties' social media posts.

Before proceeding, we reiterate that the goal of this study is not to resolve normative debates surrounding parties' multilingualism. For example, there are long-lasting debates about what kind of party system is more desirable for the stability of the state: having parties that cross-cut group boundaries and use more accommodating policies/appeals (Horowitz 1985; Reilly 2002) or having different parties that represent distinct language groups and adopt a strategy that indirectly encourages linguistic fractionalization (Laitin 1998; Lijphart 1977). Or, one might ask whether or when linguistic representation is normatively desirable, and why language-based representation in party communication matters beyond the descriptive and substantive representation of different language groups in parliament. Rather than arbitrate these claims here, our goal is to present some initial empirical patterns about what makes multilingualism in party communication more or less likely, and to provide a novel empirical tool that enables future studies to further explore language-based representation and its consequences.

## **When do parties use multilingual communication?**

Once we know which parties use multilingualism, the next question becomes what encourages them to adopt this strategy in the first place. Addressing this question is critical for two reasons. First, it allows us to understand under what conditions multilingual public spheres are likely to emerge. Second, it also sheds new light on broader debates about the relationship between social cleavages and party system (Duverger 1954; Lipset and Rokkan 1967).

We draw on a basic incentives-constraints theory of party strategy, which starts with the recognition that parties are seat maximizers and will follow an electoral strategy that will help them gain more seats. Most fundamentally, the literature assumes that social conditions define salient issues in the party system, which, in turn, influence what strategy is electorally necessary (Abou-Chadi et al. 2020; Tavits and Potter 2015; Toubeau and Wagner 2016). In the current context, this means that as linguistic diversity increases, parties should (on average) perceive greater needs to appeal to different language groups by embracing multilingualism. However, how parties respond to underlying social structures is not likely to be uniform both across and within party systems because two factors constrain what is electorally viable for them. First, institutional constraints influence parties' calculus by mechanically determining how votes are translated into seats (Calvo and Hellwig 2011; Cox 1990). Second, ideological constraints limit the range of actions parties can take without undermining their ideological brands (Adams and Somer-Topcu 2009; Lupu 2014; Tavits 2007). As we explain in greater detail below, seat-maximizing parties respond to increasing linguistic diversity and adopt multilingualism only when, within these two constraints, clear incentives exist to do so.

To begin, the institutional constraints imposed by electoral systems – how votes are translated into seats – affect the extent to which parties need to make cross-cutting multilingual appeals to win a seat in a district (Horowitz 1985; Reilly 2002). Here, we focus on the distinction between two broad categories of electoral systems: majoritarian systems, which require parties to obtain a majority of votes to gain a seat in the district, and proportional systems, which translate votes into seats in a more proportional

manner.

The two electoral systems have differential effects on how minority language groups' size is translated into their electoral strength in each district (Crisp et al. 2018; Huber 2012). Under proportional systems, the electoral strength of language minorities grows proportional to their size. The growing electoral strength of minority groups should, in turn, proportionally increase parties' willingness to use multilingual appeals.

Under majoritarian systems, however, the relationship between group size and electoral strength becomes more complex. When the size of minority language groups is very small, parties can build a majority coalition and win a seat by ignoring these groups. By contrast, when their size becomes sufficiently large, it is necessary to accommodate minority voters to build a winning coalition. This means that as their group size increases, the extent to which minority language groups' votes become pivotal grows at a much higher rate under majoritarian systems than under proportional systems (Westlake 2018). Consequently, as linguistic diversity increases, parties' incentives to use cross-cutting appeals should increase more steeply under the former than the latter (Horowitz 1985; Reilly 2002).<sup>7</sup> In sum, the constraints induced by electoral systems generate the following expectation:

**Hypothesis 1** Increases in linguistic diversity are more strongly associated with parties' adoption of multilingualism under majoritarian systems than under proportional systems.

Turning to the ideological constraints, parties' existing ideological brands restrict

---

<sup>7</sup> Electoral systems also influence the likelihood that minority language parties emerge (Bochsler 2010). It is easier to form a minority party under proportional systems than under majoritarian systems (Lijphart 1977; Norris 2008). The relative absence of minority-based parties under majoritarian systems gives additional incentives for other parties to use multilingual appeals and mobilize the untapped votes of language minorities.

whether they can use multilingual appeals to maximize their seats. Parties' choices of different appeals are electorally rewarding only if these appeals are consistent with the ideological images that parties have built up among their supporters and the electorate more generally (Adams and Somer-Topcu 2009; Lupu 2014). When voters perceive that parties deviate from their core brands, they may feel betrayed and punish parties. According to Tavits (2007), this is especially the case in the domain of principled issues, which are related to voters' core values, beliefs, and group identity.

We expect that the use of multilingualism would be detrimental to the brand maintenance of socially right-leaning parties, particularly those on the extreme right. Since their supporters tend to hold more negative views toward minority groups and cultural diversity (Golder 2016; Inglehart and Norris 2016), multilingual appeals could lead to a backlash among the core supporters of the socially rightist parties. In line with this argument, Flores and Coppock (2018) show that Spanish-language advertisements reduce candidates' electoral support among English-speaking monolingual Americans. Such backlash may be particularly likely when using minority languages on social media because, on these platforms, parties cannot choose their audience. Furthermore, prior work shows that, in anticipation of possible backlash, right-leaning parties tend to place ethnic minority candidates in lower, less visible list positions (Van der Zwan, Lubbers and Eisinga 2019). These same considerations likely incentivize socially right-leaning parties to refrain from using different languages.

In contrast, socially left-leaning parties are less likely to face the same ideological constraints. This is because they are known for their commitment to social justice and promotion of multiculturalism (Ireland 2004; Westlake 2018), and therefore communicating in different languages is unlikely to be seen as a violation of their core principles. Indeed, their supporters may even welcome parties' adoption of multilingualism as a positive signal that reinforces their ideological commitment to

cultural diversity.<sup>8</sup> In short, the ideological constraints that parties face lead to the following expectation:

**Hypothesis 2** Increases in linguistic diversity are more strongly associated with parties' adoption of multilingualism for socially left-leaning parties than for the socially right-leaning ones.

Note that this hypothesis concerns parties' left-right ideologies on the social dimension, and not on the economic one. We therefore expect that parties' economic ideologies do not predict their adoption of multilingual appeals as well.

### **Measuring parties' multilingualism using social media data**

To examine language use by parties (and candidates), we analyze messages on their official Facebook accounts. Social media data provides arguably the most appropriate tool to analyze multilingual practices for two reasons. First, it provides richer text data on day-to-day elite communication than any other source because most parties are active on social media on a near daily basis. Other text data that parties produce is ill-suited to understand their everyday language choice. For example, party websites are a more static representation of communication.<sup>9</sup> Similarly, campaign manifestos are issued only when

---

<sup>8</sup> Therefore, it is reasonable to expect that multilingual parties employ inclusive, rather than divisive tones in their communications. Analyzing the content of multilingual communication is beyond the scope of this research, but remains an important task for future studies.

<sup>9</sup> For a subset of parties in our data, we compare their multilingual practices on social media (de facto multilingualism) and websites (pro forma multilingualism). We find that the two types of multilingualism are positively correlated with  $r = 0.30$ . However, we also observe that pro forma multilingualism is more prevalent as many parties have a

there is an election. Reflecting parties' day-to-day activities, the sheer volume of text data on social media becomes incomparable to that of website or manifesto data, allowing us to have a much more accurate understanding of how parties balance communication in different languages.

Second, social media data is easier to collect than other documents that parties produce. For example, the Comparative Manifesto Project (CMP; Volkens et al. 2020) does not collect party manifestos written in second languages (if any), which makes it impossible to perform the kind of analysis we present here. Other sources of policy statements, such as news clippings, are scattered, making it hard to grasp the whole picture of parties' language use (Gadjanova 2021). By contrast, we can easily download and analyze the complete data on parties' messages on social media (assuming that we locate the relevant accounts of these parties in the first place).

In the remainder of this section, we illustrate how to use language detection tools to analyze multilingualism using parties' social media data. We first describe the data we use. Second, we explain the ideas behind the computational methods of language detection. Third, we apply these methods to label languages of individual posts on parties' Facebook pages. Finally, we develop our measure of monolingual and multilingual parties based on these post-level classifications.

### ***Party-level social media data***

We focus on parties and electoral coalitions in 87 countries. These countries were selected because they met at least one of two conditions: (1) a democratic country<sup>10</sup> included in

---

small web-section in English. This implies that relying on party websites could overestimate the extent to which they use different languages in daily interactions with voters.

<sup>10</sup> We use the average Polity scores (Marshall and Gurr 2020) between 2016 and 2018

the CMP (Volkens et al. 2020) or (2) a democratic country with a population of more than 1 million and greater than 20% Facebook penetration (Internet World Stats 2021). As a consequence of this coding rule, our dataset encompasses a diverse set of democracies from different regions, with less than one-third being advanced Western democracies.<sup>11</sup>

We collected the public Facebook pages of parties and electoral coalitions that received at least 3% of the popular vote or 1% of the seats in lower house elections that happened after 2016.<sup>12</sup> To identify the correct Facebook pages, we first checked parties' websites and obtained links to their Facebook pages. If Facebook accounts were not linked on their websites, we used search on Google and Facebook based on party names. Once we found a page, we confirmed it was a valid one by checking page description, page history, post content, and user engagements.

While some parties use multiple languages on a single page, other parties set up different pages by language.<sup>13</sup> The former practice is the norm in Canada, except for the

---

and consider a country as a democracy if its mean score is above 5.5. We supplement these cases with countries coded as electoral democracies according to Freedom House (2022) in 2022.

<sup>11</sup> The data includes 11 countries from Asia, 17 from Latin America, 4 from North Africa/the Middle East, 6 from Sub-Saharan Africa, 2 from the Caribbean, 23 from Eastern Europe, and 24 from Western Europe/North America.

<sup>12</sup> Our data collection started in 2020. For the party-level data, we also collected historical data since 2016. This was more difficult and highly labor-intensive to do for the candidate-level data, which we detail below. As a result, these data begins in 2020.

<sup>13</sup> We do not distinguish between the two types of multilingual practice as parties may switch back and forth between these types. For example, as of 2020, the Green Party of Canada had separate Facebook pages in English and French. But, as of 2021, the two

Quebec Bloc, while the latter is common in countries like Estonia, Israel, and Switzerland. In many cases, parties' websites had links to Facebook pages in different languages. However, we also conducted generic searches on Google and Facebook using party names in the country's official languages and all languages that were used in the websites of parties in the same country.<sup>14</sup> In total, we identified the official Facebook pages of more than 900 parties (93% of the parties and coalitions on our original target list).

In this study, we analyze parties' Facebook posts from 2016 to 2022, which were downloaded through Facebook's CrowdTangle API (CrowdTangle Team 2022). After excluding parties that had less than 50 posts, our dataset consists of 843 parties. In total, we analyze around 4 million posts, which together received more than 280 million user engagements.

### ***Computational language detection at scale***

Identifying the language in which texts are written poses several challenges. First, at a conceptual level, there is no universally acceptable way to define what constitutes a distinct "language" relative to a "dialect." All scholars may agree that Chinese and Spanish are distinct languages, but the lines of demarcation are often more subtle, and linguistic researchers do not always reach a consensus. Difficult examples include Croatian and Bosnian, Indonesian and Malay, and Scottish and Irish Gaelic.<sup>15</sup> Second, even given common definitions, classifying the language of any particular piece of text is

---

pages were merged into one.

<sup>14</sup> Some parties in non-English speaking countries (e.g., Lebanon and the Netherlands) establish Facebook pages in English. This is partly explained by the growing number of immigrants and diaspora voters (see, e.g., DutchNews 2022).

<sup>15</sup> For this reason, there is also no consensus on how many languages are currently spoken around the world or even in many countries.

not always straightforward. Only a handful of languages can be determined strictly by the alphabet, and the rest must be determined by words themselves. An algorithm could attempt to identify language based on the words used in a document. However, this would require training a model on a large number of words in every language in the world. Moreover, determining what constitutes a “word” is sometimes difficult without first knowing the language in which the text is written, as many languages (e.g., Chinese) do not separate words with spaces. To address these challenges, we use an ensemble of language detection algorithms. These algorithms build on interdisciplinary academic and industry research (e.g., Google) and are trained on massive amounts of data. The seven algorithms we use in this study are summarized in Table 1.

**[Table 1 about here]**

These algorithms proceed by preprocessing the text into substrings (“ $n$ -grams”), which are  $n$ -character strings from which the original text can be constructed. For example, the text string “referee” would be processed into ‘re’, ‘ef’, ‘fe’, and ‘ee’ if  $n = 2$ .<sup>16</sup> Representing text as  $n$ -grams has the useful property of not requiring any *ex ante* knowledge about the language; Chinese and English alike can be processed the same. Since there are also far fewer  $n$ -grams than words, this greatly reduces the dimensionality of the representation and enables classification even for short documents. Some approaches, including c1d3 (Ooms 2021) and fastText (Joulin, Grave, Bojanowski, Douze, Jégou and Mikolov 2016), further process the text by mapping  $n$ -grams – which are sparse features – into embeddings to further reduce the dimensionality of the features. Others, like franc (Csendi, Wormer, Ceglowski, Rideout and Johnson 2021), instead represent each document as a sparse vector of  $n$ -gram frequencies, which is

---

<sup>16</sup> Before doing so, an algorithm may first check to ensure that characters in the text do not belong to one of the few languages with a unique alphabet, but this step only identifies a tiny fraction of the world’s languages.

analogous to a document-term matrix.

After preprocessing the data, each algorithm proceeds by applying a pre-trained model to the text. The models were trained on a large corpus of documents written in a variety of languages, but where the language of origin is known. Wikipedia is commonly used since it hosts millions of documents written in dozens of languages. The algorithms we employ use a variety of models for this step. `cld2` (Ooms 2020), `langdetect` (Danilak 2021), and `langid` (Lui and Baldwin 2012) take a similar, simple approach to classification using a naive Bayes classifier on the  $n$ -gram frequencies. Other algorithms (e.g., `cld3`) employ a neural network with a large number of parameters. As part of this step, each method also limits the number of languages and/or dialects it is trained to detect. The algorithms we use detect between 56 (`langdetect`) and 206 (`franc`) languages. We provide the complete list for each model in SI Table B.1 (pp. 3-5).

### ***Post-level classification***

We asked each language detection tool to detect one language per input text. This yielded seven language labels for each post.<sup>17</sup> For the vast majority of posts, all seven methods give identical labels, while for others they disagree, mainly when one or more detection method does not cover the relevant language. Disagreements also happen when texts are very short, posts are actually written in two languages, or posts include proper nouns (names, titles, locations, etc.). As we describe in the SI Section C (pp. 6-9), we find that excluding posts with less than 125 characters reduces disagreement between methods significantly.<sup>18</sup> Thus, in the subsequent analyses, we use the 125 characters threshold.

---

<sup>17</sup> Before doing so, we did simple preprocessing of post texts (lower-casing letters and removing URLs, emoji, punctuation, hashtags, usernames, and numbers).

<sup>18</sup> Once we exclude the posts with less than 125 characters, they give the identical language label for 75.7% of the posts, and at least four methods give the same label for

In three cases, we merged languages to further improve performance. First, we treat Indonesian and Malay as a single language (Indonesian-Malay). This is because the language detection methods provided inconsistent labeling for posts by Indonesian and Malaysian parties. Since the two languages are mutually intelligible (Adelaar and Himmelmann 2004; Wichmann 2020), and we can safely assume that Indonesian (Malaysian) parties do not use Malay (Indonesian), collapsing them as a single language gives more reliable classification results. As we detail in the SI Section D (pp. 10-11), human coding of Indonesian and Malaysian parties' posts validates this decision. Second, we found that the method had difficulty distinguishing Central Kurdish and Persian in several Iraqi parties. Since human coding showed that these posts were actually Central Kurdish and we have no data from Iran itself, we simply collapse these languages into Central Kurdish-Persian.<sup>19</sup>

A remaining concern are the posts from Bosnia and Herzegovina, Croatia, Montenegro, and Serbia, where the primary languages are Bosnian, Croatian, Montenegrin, and Serbian (if you view them as separate) or Serbo-Croatian (if you view them all as one language). These “politically divorced” (Laitin 2000) languages are particularly difficult cases since they are mutually intelligible and share a common grammar and spelling, making them difficult to distinguish from texts alone. This means that we observed high rates of post-level disagreements between methods.<sup>20</sup> In our main results, we choose to combined them into one Serbo-Croatian because, “[w]hat is clear to everyone [...] is that all of these languages share a common core, a fact which enables

---

98.5% of the posts. More than 60% of the posts have more than 125 characters.

<sup>19</sup> Additional details of the human coding are provided in the SI Section C (pp. 6-9).

<sup>20</sup> Indeed, if we treat these languages as separate ones, many parties in the region are labeled as multilingual, even in cases where this makes little sense. For instance, nearly half of posts in Croatia were coded as Bosnian by many methods even though Bosniaks make up less than 1% of the population.

all their speakers to communicate freely with one another” (Alexander 2006, p. xvii). To the extent that our argument rests on concerns about creating a common public sphere, this means that while language choice might serve as an important identity signal to voters, it does not impede actual communication and dialogue *per se* in this setting. However, in the SI Section E (pp. 12-14), we also report results where, first, they are treated as separate, and, second, where these cases are removed entirely. All results are essentially unchanged.

To understand what languages appear in our party-level corpus, we explore the relationship between the proportion of first-language (L1) users and that of parties’ Facebook posts written in the corresponding languages. The former is based on *Ethnologue* (Eberhard, Simons and Fennig 2022),<sup>21</sup> and we focus on the three most spoken languages in each country that are used by more than 5% of the population. We also exclude languages only two or fewer detection methods can detect. To measure the proportion of each language in party communication, we take the average proportion of the posts written in that language across detection methods (if they include the language).<sup>22</sup>

Note that using the census-based proportion of L1 users as a metric of a language community size raises several conceptual issues. To begin, first language maps strongly onto ethnicity and may not fully capture the actual communicative practice of the country, such as the presence of a *lingua franca* or the possibility that people with different first languages can communicate with each other using a shared secondary language (Laitin 2000; Liu and Pizzi 2018). For some of the language groups in our sample (e.g., Guarani in Paraguay and Zulu in South Africa), the number of L1 users is smaller than the total number of speakers who can communicate using another language (e.g.,

---

<sup>21</sup> See <https://www.ethnologue.com/>.

<sup>22</sup> For example, only five detection methods can detect Armenian. Hence, the estimated proportion of Armenian posts among Armenian parties is computed on these five methods.

more people overall speak Spanish than Guarani in Paraguay, even though Guarani has the largest community as measured by L1). In addition, it is not always clear how people interpret language questions in a census. Some scholars even suggest that choosing a certain native language in the census captures an expression of people's loyalty to an ethnic group and broader political preferences, rather than their communicative competence (Arel 2002; Kulyk 2018). Consequently, the number of L1 users may reflect not just how many people use the language but also how many people strongly identify themselves with that language and care about it (Kulyk 2011).

With these limitations in mind, Figure 1 summarizes our exploratory findings. The horizontal axis shows the proportion of L1 users of a given language-country, and the vertical axis is the estimated proportion of parties' Facebook posts written in the same language aggregated at the level of the country. Panels are separated by the rank of language groups based on their population shares. The dashed lines indicate 45-degree lines. Language groups located above (below) these lines mean that their languages are overrepresented (underrepresented) in party communication relative to their population shares. The solid lines show fitted linear models based on regressing the two axes.

**[Figure 1 about here]**

In general, we find a strong positive association between the population share of the language and the extent to which it appears in party communication, as indicated by positive fitted lines. This gives first-step validation to our post-level classification because parties should be more likely to use a language that is more common in the population. But more interestingly, these estimated linear lines nearly overlap the 45-degree lines. This means that there is a clear correspondence between the proportion of the language spoken in the country and that of the posts written in the same language. Hence, party communication on social media in the *aggregate* seems to give proportional attention to different language groups.

Additionally, several of the language-country pairs that most strongly diverge from

the 45-degree lines validate our post-level classification because they all represent cases in which the proportion of L1 users is arguably a poor proxy for the size of the actual language community. This includes Zulu in South Africa in Panel (a), Indonesian and Javanese in Indonesia in Panels (a) and (b), Spanish in Paraguay in Panel (b), and Russian in Kyrgyzstan in Panel (c). All of these are cases where one language (English, Indonesian, Spanish, and Russian) serves as a *lingua franca* for political communication even where it is not the largest L1 language.

### ***Party-level classification***

Using the post-level languages labels, the final step is to construct a party-level measure of multilingualism. The difficulty here is distinguishing between cases where parties truly include texts in multiple languages and cases where we are simply mis-measuring language use at the post level. This is particularly important for languages not included in some or all of the language classifiers.

To mitigate these concerns, we do **not** use the specific language the algorithms identify. Instead, we focus on a dichotomous measure of whether the party is primarily multilingual or monolingual. The advantage of this approach is that we do not need to identify the “correct” language for each post or the overall proportion of posts associated with each language. Our findings above suggest that existing language classifiers are simply too inaccurate to provide reliable estimates for these quantities at scale. By instead focusing on multilingualism, we only require that each method consistently assigns the *same* label to any one language (and different labels to different languages) rather than the *correct* label. For example, `langdetect` and `tika` do not include Armenian and give an Estonian label to posts by Armenian parties. However, since this “wrong” labeling is internally consistent, it does not impact the final multilingual classification of Armenian parties.

Specifically, we classify parties as monolingual or multilingual using the following steps. First, for each detection method, we labeled the party as primarily multilingual if

the proportion of the most detected language is below 90% (meaning more than 10% of posts are not from the dominant language). Second, we created a party  $\times$  detection method matrix  $\mathbf{Y}_{p \times j}$ , where  $Y_{pj} = 1$  if party  $p$  is multilingual according to detection method  $j$ , and  $Y_{pj} = 0$  otherwise. Third, we fit a latent class model (Linzer and Lewis 2011) to synthesize these judgements.<sup>23</sup>

In addition, we hired human coders to label the language of 300 randomly selected posts for 44 difficult cases where either our labels showed high levels of disagreement between methods or where we were not detecting languages widely spoken in a country (suggesting we may be missing a language).<sup>24</sup> In only five cases, the human coders significantly disagreed with the latent class labels. Four were parties in Timor-Leste, which resulted from poor handling of Tetum. The fifth was an Iraqi party where the methods struggled with Central Kurdish. In these five cases, we used the human coder judgements instead.

In the end, we classify 101 parties (12%) as primarily multilingual.<sup>25</sup> We find multilingual parties in 28 countries. Of these, 25 countries have both monolingual and multilingual parties, while in Lesotho, Kyrgyzstan, and Mauritius, all parties practice multilingualism. In the remaining 59 countries, parties are all classified as primarily monolingual. These cases include linguistically homogeneous countries and those where national politics operates in a single language despite the presence of various language communities. In most of these 59 countries, all parties use the same language. However, in Kosovo, Romania, and Ukraine, monolingual parties use different languages to appeal

---

<sup>23</sup> Additional details of these steps are provided in the SI Section F (pp. 15-17).

<sup>24</sup> Specifically, we examined eight parties in Botswana, Kyrgyzstan, Luxembourg, Nepal, New Zealand, North Macedonia, and Romania. We also re-examined all parties in Indonesia, Iraq, Lesotho, Malaysia, and Timor-Leste. Additional details for this exercise are provided in SI Sections C-F (pp. 6-17).

<sup>25</sup> See SI Tables G.1 and G.2 (pp. 18-20).

to distinct language groups. Put differently, these countries represent cases in which the party systems show a complete split along linguistic lines.

To validate our measure, we follow Adcock and Collier (2001) and inspect convergent and divergent cases by analyzing a relationship between linguistic diversity and the proportion of primarily multilingual parties by country. The measure of linguistic diversity is based on ethno-linguistic fractionalization (ELF) by Desmet et al. (2012) because it covers the most countries, with Kosovo as the only missing case. They provide 15 different fractionalization indices at different levels of group aggregation based on language trees from *Ethnologue*. We use their measure at the second lowest level of aggregation, *ELF* (14), as it is most highly correlated with other widely used measures of ELF (Alesina, Devleeschauwer, Easterly, Kurlat and Wacziarg 2003; Desmet, Weber and Ortuño-Ortín 2009).

Since the measure of ELF relies on census data, the conceptual concerns about L1 users that we discussed above also apply to this measure. That is, it may overestimate linguistic diversity in countries with a *lingua franca* and capture not so much people's actual communicative practices as the intensity of their identification with certain ethnic/language groups (Kulyk 2011; Laitin 2000).<sup>26</sup> Nevertheless, we think that this is a reasonable statistic in the current context because more groups with different languages mean that there are more languages in the country.

Figure 2 shows the relationship between linguistic diversity and the proportion of multilingual parties. The solid linear line indicates that there is a positive association between the two. This makes sense because seat-maximizing parties should have greater incentives to use multilingual appeals as linguistic diversity increases. Moreover, we find no multilingual party in countries like Indonesia, Namibia, and Senegal, where, despite high scores on linguistic diversity, a single language is used as the medium of

---

<sup>26</sup> Laitin (2000) also notes that *Ethnologue's* language trees are not equally sensitive to dialectical differences across regions.

government. This provides additional confidence to our classification results.

**[Figure 2 about here]**

## **Analyses**

Hypotheses 1 and 2 suggest that institutional and ideological constraints moderate the relationship between linguistic diversity and parties' adoption of multilingualism. The sources and operationalization of the key variables are as follows.

We use two country-level variables: linguistic diversity and electoral systems. As we explained above, the former is based on Desmet et al. (2012). For electoral systems, we rely on V-Dem (Coppedge et al. 2021) and assign the value of 1 for countries using majoritarian systems and 0 otherwise. During the period under study, Mongolia switched from a non-majoritarian system to a majoritarian one. Since we do not analyze the temporal dynamics of parties' multilingualism in this study, in the case of Mongolia, we use the modal system between 2016 and 2022. Thirteen countries are coded as majoritarian systems.

To capture parties' ideological orientations, we rely on the Global Party Survey (GPS; Norris 2019). The advantage of the GPS is that it provides two separate measures for parties' social/cultural and economic ideologies.<sup>27</sup> We expect that social/cultural left ideology, rather than economic left ideology, drives parties' multilingualism. For both measures, 0 indicates that parties are extreme left, whereas 10 means that they are extreme right. The correlation between the two variables is only moderate ( $r = 0.46$ ).<sup>28</sup>

We fit a linear probability model with linguistic diversity, a measure for our

---

<sup>27</sup> The GPS does not include Honduras, Kosovo, Senegal, and Sri Lanka. Also, it does not cover small or relatively new parties.

<sup>28</sup> Descriptive statistics are in SI Table H.1 (p. 21).

hypothesized moderator, and the interaction of the two.<sup>29</sup> Our unit of analysis is the party, and the outcome is a dummy indicator of primarily multilingual parties. We also weight observations by the logged number of posts. In Models 1-3 of Table 2, we consider each moderating variable separately, which has the advantage of maximizing the number of parties we can include to test each hypothesis. In Models 4 and 5, however, we combine them into a single model specification and ensure that our results remain the same.

**[Table 2 about here]**

### ***H1: Electoral system***

We begin by analyzing the moderating roles of electoral systems. Panel (a) of Figure 3 summarizes the marginal effect of linguistic diversity on parties' adoption of multilingualism under different electoral systems, based on Model 1 of Table 2. Vertical bars indicating 95% confidence intervals. We find that under proportional systems, the coefficient on linguistic diversity is 0.42 with a 95% confidence interval of [0.18, 0.65]. By contrast, under majoritarian systems, the coefficient is 0.76 with a 95% confidence interval of [0.56, 0.95].

**[Figure 3 about here]**

The difference between the two estimates (the interaction term  $ELF \times$  majoritarian system) is 0.34 and statistically significant with a 95% confidence interval of [0.04, 0.65]. Further, this difference is substantively meaningful as switching from non-majoritarian to majoritarian systems can increase the marginal effect of linguistic diversity by 81%.

---

<sup>29</sup> We do not use logistic regression because there is virtually no multilingual party when linguistic diversity is low (especially under majoritarian systems), meaning that our data has a problem of quasi-complete separation. Using logistic regression would lead to convergence failures.

Therefore, our results are generally consistent with the notion that the relationship between linguistic diversity and parties' adoption of multilingualism is more pronounced under majoritarian rules.

## ***H2: Party ideology***

Next, we analyze the moderating roles of party ideology in Models 2 and 3 of Table 2.<sup>30</sup> We find that the interaction term  $\text{ELF} \times \text{social ideology}$  is negative and statistically different from 0 with a 95% confidence interval of  $[-0.14, -0.01]$ . By contrast, the interaction terms  $\text{ELF} \times \text{economic ideology}$  does not reach the conventional levels of statistical significance with a 95% confidence interval of  $[-0.10, 0.02]$ . These results suggest that parties' left-right ideology in the social dimension can condition their multilingualism, but their ideology in the economic dimension may not.

Panel (b) of Figure 3 shows the marginal effect of linguistic diversity on parties' multilingualism conditional on social/cultural ideology, based on Model 2 of Table 2. It shows that socially/culturally left-leaning parties are more likely to translate linguistic diversity into multilingualism than their right-leaning counterparts. Indeed, for extreme right parties, the marginal effect of linguistic diversity becomes statistically insignificant. These patterns are consistent with the argument that socially leftist parties become more willing to accommodate different language groups as linguistic diversity increases, while socially right-leaning parties are unresponsive to changes in linguistic diversity.

## **Extension: candidates' language use**

In this section, we analyze candidate-level language practices by exploring whether parties' multilingualism is reflected in the language use of candidates they nominate.

---

<sup>30</sup> We do not find that the moderating roles of ideology deviate from linearity (Hainmueller, Mummolo and Xu 2019).

Addressing this question is important because parties not only influence which languages are used in democratic deliberation but also play a gatekeeper role in recruiting candidates from different language backgrounds (Dancygier, Lindgren, Nyman and Vernby 2021; Eriksson and Vernby 2021).

We expect that parties' multilingualism is strongly correlated with the language practice of the candidates they recruit. If so, candidates nominated by multilingual parties tend to adopt linguistic appeals similar to the ones used by their parties. This implies either that candidates for multilingual parties are more likely to use multilingual appeals than those for monolingual parties or that the former are more willing to use a non-dominant language than the latter. Although these expectations are intuitive, evaluating these propositions is still theoretically valuable. After all, it allows us to understand whether multilingual communication is merely a cheap talk/symbolic gesture by parties to woo different language groups (Chandra 2011; Devasher and Gadjanova 2021) or it is translated into actual multilingual communication by candidates.

We analyze candidate-level Facebook data collected in twelve lower house elections that happened between December 2020 and April 2023: Canada (2021), Cyprus (2021), Denmark (2022), Estonia (2023), Finland (2023), Israel (2021), Latvia (2022), Malaysia (2022), the Netherlands (2021), the Philippines (2022), Romania (2020), and Serbia (2022). We selected these countries because they show within-country variation in the presence of multilingual parties and the languages that parties use.<sup>31</sup> Before each of these elections, we obtained the list of candidates from the website of the election commission. Then, performing the same generic searches as those for parties, we collected the public Facebook pages of candidates. The coverage of Facebook accounts among candidates

---

<sup>31</sup> Every country but Romania has at least one multilingual party. Romania is an unusual case where multiple languages are used, but all parties were categorized as monolingual reflecting a clean partition of parties by language.

varies by country,<sup>32</sup> and our analysis is restricted to those who used public Facebook pages.

Using CrowdTangle API, we downloaded these candidates' posts within the 60-day window of the elections. We focus only on candidates running for the parties included in the party-level data. We also exclude candidates with fewer than 10 posts. This leaves 4,188 candidates for 111 parties. The total number of posts we analyze exceeds 37,000.

We relied on the same procedures as the ones for the party-level classification to measure candidates' language practices. The result is a dummy variable that equals one for candidates engaged in multilingual communication. We also created a dummy indicator of whether the most used language by the candidate is a minority language of the country. As SI Table I.1 (p. 22) summarizes, 417 candidates (10.0%) are primarily multilingual, and 433 candidates (10.3%) mainly communicate in a language other than the most dominant one in their countries.

The unit of analysis is candidate  $i$  in party  $p$  in country  $c$ . The outcomes are whether the candidate is primarily multilingual and whether the candidate mainly uses a minority language. The key predictor is a dummy indicator of whether the candidate's party is primarily multilingual. We control for parties' social ideology as well as countries' electoral systems and linguistic diversity. The model specification is based on a multilevel linear probability model with nested random effects by party and country. We weight observations by the logged number of posts.

Table 3 summarizes the results. In Models 1 and 2, the outcome is a dummy indicator of whether candidates are primarily multilingual. The estimates of primarily multilingual party are positive and statistically discernible from 0 ( $p < 0.01$ ), regardless

---

<sup>32</sup> While more than 75% of the candidates in Canada set up public Facebook pages, this number goes down to around 20% in Israel, the Netherlands, and Serbia. Candidates under candidate-centric electoral systems are more likely to use a public Facebook page than those under party-centric systems.

of whether we control for the key party- and country-level covariates. This means that the candidates of primarily multilingual parties are more likely to embrace multilingualism than those of monolingual parties.

### **[Table 3 about here]**

Next, Models 3 and 4 of Table 3 analyze whether candidates mainly communicate in a minority language. The estimates of primarily multilingual party are positive and statistically discernible from 0, although only at the 0.10 level in Model 4. This means that party-level multilingualism is positively correlated not only with candidate-level multilingualism but also with how much candidates are willing to use a non-majority language.

## **Conclusion**

This study analyzes parties' and candidates' Facebook posts to understand their day-to-day language practices. More specifically, by applying computational tools of language detection, we generate the first classification of primarily monolingual and multilingual parties in 87 democracies. The resulting dataset provides the most comprehensive descriptive account of which parties embrace multilingualism to date.

Substantively, this study offers two novel insights into the linguistic aspect of party competition. First, we show that the interplay between countries' linguistic composition and parties' electoral incentives is related to their adoption of multilingual appeals. Specifically, majoritarian systems tend to encourage parties to cut across language lines more than non-majoritarian systems when linguistic diversity is high. Further, socially/culturally left-leaning parties are more likely to adopt multilingualism than right-leaning parties as linguistic diversity increases. These findings deepen our understanding of the strategic behavior of seat-maximizing parties. Second, extending our analysis to the language use of candidates in a dozen of multilingual democracies,

we show that candidates nominated by multilingual parties tend to adopt similar multilingual communication strategies to those of their parties.

The implications of these findings are important. They indicate that the presence of multilingual parties matters not just because they can communicate with different language groups. Rather, their presence has far-reaching consequences on how language spaces are structured by individual candidates during elections. Because they can effectively dampen, rather than reinforce, language-based divisions, multilingual parties and candidates may transform representation from being group-based to being interest-based. This could in theory decrease the level of political conflict and instability.

The dataset we provide opens up various avenues for future research. First, this study is by its nature descriptive, providing a cross-sectional snapshot of party and candidate behavior. The correlations we report are consistent with our theory, but future work may seek to provide a stronger ground for establishing causal claims. For instance, as we accumulate more data over time from social media, scholars might investigate *changing* behavior in response to electoral reforms.

Second, it is critical to assess overtime changes in parties' multilingualism. It is especially interesting to examine how election timing influences parties' decisions to use multilingual appeals, as some parties may try to accommodate minority voters only when elections get closer. Third, it is important to ask what additional factors determine parties' adoption of multilingualism. Although we explore two key determinants, there is still room for further theorization and empirical evaluation. We also envision examining parties' decisions to use different languages at the level of the post, which requires analyzing the post content. Finally, we should examine how multilingual parties influence voter behavior. By combining our dataset and survey data, we can analyze how multilingual parties shape voters' political attitudes, such as institutional trust and perceptions of parties. All these questions will eventually help us understand the overall impact of multilingual parties on the quality of democratic deliberation.

As a final note, we caution downstream researchers against blindly using the

language labels of the posts or parties, rather than the dichotomous multilingualism labels, in their applications. As we discussed above, existing language detection tools do not always give the “correct” language labels to the texts from some countries. What we propose in this study is an approach to classify parties into monolingual and multilingual ones that is plausibly resilient to this problem. If researchers want to use a specific language label as an explanatory or outcome variable, it is important to perform additional steps to validate detected languages or pick the right computational tool of language detection that is suited for a specific case. We highlight cases where this issue may be specifically important in the SI Section F (pp. 15-17).

## References

- Abou-Chadi, Tarik, Christoffer Green-Pedersen and Peter B Mortensen. 2020. "Parties' Policy Adjustments in Response to Changes in Issue Saliency." *West European Politics* 43(4):749-771.
- Adams, James and Zeynep Somer-Topcu. 2009. "Policy Adjustment by Parties in Response to Rival Parties' Policy Shifts: Spatial Theory and the Dynamics of Party Competition in Twenty-Five Post-War Democracies." *British Journal of Political Science* 39(4):825-846.
- Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529-546.
- Addis, Adeno. 2007. "Constitutionalizing Deliberative Democracy in Multilingual Societies." *Berkeley Journal of International Law* 25(2):117-164.
- Adelaar, K Alexander and Nikolaus Himmelmann. 2004. *The Austronesian Languages of Asia and Madagascar*. New York: Routledge.
- Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8(2):155-194.
- Alexander, Ronelle. 2006. *Bosnian, Croatian, Serbian: A Grammar with Sociological Commentary*. Madison, WI: University of Wisconsin Press.
- Arel, Dominique. 2002. "Interpreting 'Nationality' and 'Language' in the 2001 Ukrainian Census." *Post-Soviet Affairs* 18(3):213-249.
- Barrington, Lowell. 2022. "A New Look at Region, Language, Ethnicity and Civic National Identity in Ukraine." *Europe-Asia Studies* 74(3):360-381.

- Bochsler, Daniel. 2010. "Electoral Rules and the Representation of Ethnic Minorities in Post-Communist Democracies." *European Yearbook of Minority Issues* 7(1):153-180.
- Bonotti, Matteo and Nenad Stojanović. 2022. "Multilingual Parties and the Ethics of Partisanship." *The Journal of Politics* 84(1):470-482.
- Caluwaerts, Didier and Min Reuchamps. 2014. Deliberative Stress in Linguistically Divided Belgium. In *Democratic Deliberation in Deeply Divided Societies: From Conflict to Common Ground*, ed. Juan E. Ugarriza and Didier Caluwaerts. New York: Springer pp. 35-52.
- Calvo, Ernesto and Timothy Hellwig. 2011. "Centripetal and Centrifugal Incentives under Different Electoral Systems." *American Journal of Political Science* 55(1):27-41.
- Chandra, Kanchan. 2007. *Why Ethnic Parties Succeed: Patronage and Ethnic Head Counts in India*. New York: Cambridge University Press.
- Chandra, Kanchan. 2011. "What Is an Ethnic Party?" *Party Politics* 17(2):151-169.
- Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I Lindberg, Jan Teorell, Nazifa Alizada, David Altman, Michael Bernhard, Agnes Cornell, M Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Allen Hicken, Garry Hindle, Nina Ilchenko, Joshua Krusell, Anna Lührmann, Seraphine F Maerz, Kyle L Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Yi-ting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2021. "V-Dem [Country-Year/Country-Date] Dataset v11.1.". Varieties of Democracy (V-Dem) Project. <https://doi.org/10.23696/vdemds21>.

- Cox, Gary W. 1990. "Centripetal and Centrifugal Incentives in Electoral Systems." *American Journal of Political Science* 34(4):903–935.
- Crisp, Brian F, Betül Demirkaya, Leslie A Schwindt-Bayer and Courtney Millian. 2018. "The Role of Rules in Representation: Group Membership and Electoral Incentives." *British Journal of Political Science* 48(1):47–67.
- CrowdTangle Team. 2022. CrowdTangle. Facebook, Menlo Park, California, United States.  
List ID: GlobalPartyList.
- Csardi, Gabor, Titus Wormer, Maciej Ceglowski, Jacob R Rideout and Kent S Johnson. 2021. "franc: Detect the Language of Text." *R Package Version 1.1.4*.
- Dancygier, Rafaela, Karl-Oskar Lindgren, Pär Nyman and Kåre Vernby. 2021. "Candidate Supply is Not a Barrier to Immigrant Representation: A Case-Control Study." *American Journal of Political Science* 65(3):683–698.
- Danilak, Michal Mimino. 2021. "langdetect." *Python Module Version 1.0.9*.
- De Bres, Julia, Gabriel Rivera Cosme and Angela Remesch. 2020. "Walking the Tightrope of Linguistic Nationalism in a Multilingual State: Constructing Language in Political Party Programmes in Luxembourg" *Journal of Multilingual and Multicultural Development* 41(9):779–793.
- Desmet, Klaus, Ignacio Ortuño-Ortín and Romain Wacziarg. 2012. "The Political Economy of Linguistic Cleavages." *Journal of Development Economics* 97(2):322–338.
- Desmet, Klaus, Shlomo Weber and Ignacio Ortuño-Ortín. 2009. "Linguistic Diversity and Redistribution." *Journal of the European Economic Association* 7(6):1291–1318.
- Devasher, Madhavi and Elena Gadjanova. 2021. "Cross-Ethnic Appeals in Plural Democracies." *Nations and Nationalism* 27(3):673–689.

- DutchNews. 2022. "Local Elections: Around Half the Parties Have Info in English." *DutchNews*. URL: <https://www.dutchnews.nl/2022/03/local-elections-around-half-the-parties-have-info-in-english/>.
- Duverger, Maurice. 1954. *Political Parties: Their Organization and Activity in the Modern State*. New York: Methuen & Co. Ltd.
- Eberhard, David M, Gary F Simons and Charles D Fennig. 2022. *Ethnologue: Languages of the World*. Twenty-Fifth Edition. Dallas, Texas: SIL International. <http://www.ethnologue.com>.
- Eriksson, Lina M and Kåre Vernby. 2021. "Welcome to the Party? Ethnicity and the Interaction between Potential Activists and Party Gatekeepers." *The Journal of Politics* 83(4):1861-1866.
- Flores, Alejandro and Alexander Coppock. 2018. "Do Bilinguals Respond More Favorably to Candidate Advertisements in English or in Spanish?" *Political Communication* 35(4):612- 633.
- Freedom House. 2022. *Freedom in the World 2022*. Washington, DC: Freedom House.
- Frye, Timothy. 2015. "What Do Voters in Ukraine Want?: A Survey Experiment on Candidate Ethnicity, Language, and Policy Orientation." *Problems of Post-Communism* 62(5):247-257.
- Gadjanova, Elena. 2021. "Status-Quo or Grievance Coalitions: The Logic of Cross-Ethnic Campaign Appeals in Africa's Highly Diverse States." *Comparative Political Studies* 54(3-4):652-685.
- Golder, Matt. 2016. "Far Right Parties in Europe." *Annual Review of Political Science* 19:477-497.
- Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve

- Empirical Practice." *Political Analysis* 27(2):163–192.
- Horowitz, Donald L. 1985. *Ethnic Groups in Conflict*. Berkeley, CA: University of California Press.
- Huber, John D. 2012. "Measuring Ethnic Voting: Do Proportional Electoral Laws Politicize Ethnicity?" *American Journal of Political Science* 56(4):986–1001.
- Inglehart, Ronald F and Pippa Norris. 2016. "Trump, Brexit, and the rise of populism: Economic have-nots and cultural backlash." HKS Working Paper No. RWP16-026.
- Internet World Stats. 2021. Internet World Stats: Usage and Population Statistics. <https://www.internetworldstats.com/stats2.htm>.
- Ireland, Patrick Richard. 2004. *Becoming Europe: Immigration, Integration, and the Welfare State*. Pittsburgh, PA: University of Pittsburgh Press.
- Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou and Tomas Mikolov. 2016. "FastText.zip: Compressing Text Classification Models." *arXiv preprint arXiv:1612.03651*.
- Kulyk, Volodymyr. 2011. "Language Identity, Linguistic Diversity and Political Cleavages: Evidence from Ukraine." *Nations and Nationalism* 17(3):627–648.
- Kulyk, Volodymyr. 2018. "Shedding Russianness, Recasting Ukrainianness: The Post-Euromaidan Dynamics of Ethnonational Identifications in Ukraine." *Post-Soviet Affairs* 34(2-3):119–138.
- Kumove, Michael. 2022. "Does Language Foster Reconciliation? Evidence from the Former Yugoslavia." *Journal of Conflict Resolution* 66(4-5):783–808.
- Lacey, Joseph. 2017. *Centripetal Democracy: Democratic Legitimacy and Political Identity in Belgium, Switzerland, and the European Union*. New York: Oxford University Press.

- Laitin, David D. 1989. "Linguistic Revival: Politics and Culture in Catalonia." *Comparative Studies in Society and History* 31(2):297-317.
- Laitin, David D. 1998. *Identity in Formation: The Russian-Speaking Populations in the Near Abroad*. Ithaca, New York: Cornell University Press.
- Laitin, David D. 2000. "What Is a Language Community?" *American Journal of Political Science* 44(1):142-155.
- Lee, Taeku and Efrén O Pérez. 2014. "The Persistent Connection between Language-of-Interview and Latino Political Opinion." *Political Behavior* 36:401-425.
- Lijphart, Arend. 1977. *Democracy in Plural Societies: A Comparative Exploration*. New Haven, CT: Yale University Press.
- Linzer, Drew A and Jeffrey B Lewis. 2011. "poLCA: An R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software* 42(10):1-29.
- Lipset, Seymour and Stein Rokkan. 1967. Cleavage Structures, Party Systems, and Voter Alignments: An Introduction. In *Party Systems and Voter Alignments: Cross-National Perspectives*, ed. Seymour Lipset and Stein Rokkan. New York: Free Press pp. 1-64.
- Liu, Amy H. 2015. *Standardizing Diversity: The Political Economy of Language Regimes*. Philadelphia, PA: University of Pennsylvania Press.
- Liu, Amy H and Elise Pizzi. 2018. "The Language of Economic Growth: A New Measure of Linguistic Heterogeneity." *British Journal of Political Science* 48(4):953-980.
- Lui, Marco and Timothy Baldwin. 2012. langid.py: An Off-the-Shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*. pp. 25-30.

- Lupu, Noam. 2014. "Brand Dilution and the Breakdown of Political Parties in Latin America." *World Politics* 66(4):561–602.
- Marquardt, Kyle L. 2022. "Language, Ethnicity and Separatism: Survey Results from Two Post-Soviet Regions." *British Journal of Political Science* 52(4):1831–1851.
- Marshall, Monty G and Ted Robert Gurr. 2020. Polity5 Annual Time-Series, 1946-2018. The Polity V Project. Center for Systemic Peace. <https://www.systemicpeace.org>.
- Medeiros, Mike. 2017. "Refining the Influence of Language on National Attachment: Exploring Linguistic Threat Perceptions in Quebec." *Nationalism and Ethnic Politics* 23(4):375– 390.
- Norris, Pippa. 2008. *Driving Democracy: Do Power-Sharing Institutions Work*. New York: Cambridge University Press.
- Norris, Pippa. 2019. The Global Party Survey, 2019. V1.0. <https://www.GlobalPartySurvey.org>.
- Ooms, Jeroen. 2020. "cld2: Google's Compact Language Detector 2." *R Package Version 1.2.1*.
- Ooms, Jeroen. 2021. "cld3: Google's Compact Language Detector 3." *R Package Version 1.4.2*.
- Pérez, Efrén and Margit Tavits. 2022. *Voicing Politics: How Language Shapes Public Opinion*. Princeton, NJ: Princeton University Press.
- Pérez, Efrén O and Margit Tavits. 2019. "Language heightens the political salience of ethnic divisions." *Journal of Experimental Political Science* 6(2):131–140.
- Reilly, Ben. 2002. "Electoral Systems for Divided Societies." *Journal of Democracy* 13(2):156–170.

- Ricks, Jacob I. 2020. "The Effect of Language on Political Appeal: Results from a Survey Experiment in Thailand." *Political Behavior* 42(1):83–104.
- Ringe, Nils. 2022. *The Language(s) of Politics: Multilingual Policy-Making in the European Union*. Ann Arbor, MI: University of Michigan Press.
- Rubin, Aviad. 2014. Language Policy and Inter-Group Deliberation in Israel. In *Democratic Deliberation in Deeply Divided Societies: From Conflict to Common Ground*, ed. Juan E. Ugarriza and Didier Caluwaerts. New York: Springer pp. 151–171.
- Schmidt, Ronald. 2014. "Democratic Theory and the Challenge of Linguistic Diversity." *Language Policy* 13:395–411.
- Stojanović, Nenad and Matteo Bonotti. 2020. "Political Parties in Deeply Multilingual Polities: Institutional Conditions and Lessons for the EU." *JCMS: Journal of Common Market Studies* 58(3):599–615.
- Strani, Katerina. 2020. Multilingualism and Politics Revisited: The State of the Art. In *Multilingualism and Politics: Revisiting Multilingual Citizenship*, ed. Katerina Strani. New York: Springer pp. 17–45.
- Strijbis, Oliver and Michal Kotnarowski. 2015. "Measuring the Electoral Mobilization of Ethnic Parties: Towards Comparable Indicators." *Party Politics* 21(3):456–469.
- Tavits, Margit. 2007. "Principle vs. Pragmatism: Policy Shifts and Political Competition." *American Journal of Political Science* 51(1):151–165.
- Tavits, Margit and Joshua D Potter. 2015. "The Effect of Inequality and Social Identity on Party Strategies." *American Journal of Political Science* 59(3):744–758.
- Toubeau, Simon and Markus Wagner. 2016. "Party Competition over Decentralisation: The Influence of Ideology and Electoral Incentives on Issue Emphasis." *European Journal of Political Research* 55(2):340–357.

- Van der Zwan, Roos, Marcel Lubbers and Rob Eisinga. 2019. "The Political Representation of Ethnic Minorities in the Netherlands: Ethnic Minority Candidates and the Role of Party Characteristics." *Acta Politica* 54(2):245–267.
- Volgens, Andrea, Tobias Burst, Werner Krause, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, Bernhard Weßels and Lisa Zehnter. 2020. The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2020a. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). <https://doi.org/10.25522/manifesto.mpbs.2020a>.
- Westlake, Daniel. 2018. "Multiculturalism, Political Parties, and the Conflicting Pressures of Ethnic Minorities and Far-Right Parties." *Party Politics* 24(4):421–433.
- Wichmann, Søren. 2020. "How to Distinguish Languages and Dialects." *Computational Linguistics* 45(4):823–831.
- Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton, NJ: Princeton University Press.
- Zárate, Marques G, Enrique Quezada-Llanes and Angel D Armenta. 2024. "Se Habla Español: Spanish-Language Appeals and Candidate Evaluations in the United States." *American Political Science Review* 118(1):363–379.

Table 1: Summary of seven language detection methods

	cld2	cld3	fastText	franc	langdetect	langid	tika
Classifier	NB	NN	NN	RB	NB	NB	NB or NN
# of languages	88	102	128	206	56	95	71

*Note:* Column headings list each language detection method we use in this study, and table entries specify for each, which classifier it uses and how many languages it covers. ‘NB’ indicates that an algorithm used a naive Bayes classifier, ‘NN’ a neural network, ‘RB’ a rule-based classifier (e.g., a dictionary).

Table 2: Electoral systems, party ideology, and parties' use of multilingualism

	Primarily multilingual party				
	(1)	(2)	(3)	(4)	(5)
ELF	0.42** (0.12)	0.90** (0.23)	0.67** (0.20)	0.79** (0.23)	0.60** (0.19)
Majoritarian system	-0.09 (0.06)			-0.14* (0.05)	-0.15* (0.06)
ELF × majoritarian system	0.34* (0.15)			0.53** (0.15)	0.55** (0.17)
Social ideology		0.00 (0.01)		0.00 (0.01)	
ELF × social ideology		-0.07* (0.03)		-0.07* (0.03)	
Economic ideology			0.00 (0.01)		0.00 (0.01)
ELF × economic ideology			-0.04 (0.03)		-0.05 (0.03)
N of parties	835	444	440	444	440
N of countries	86	83	83	83	83
R <sup>2</sup>	0.15	0.19	0.15	0.22	0.18

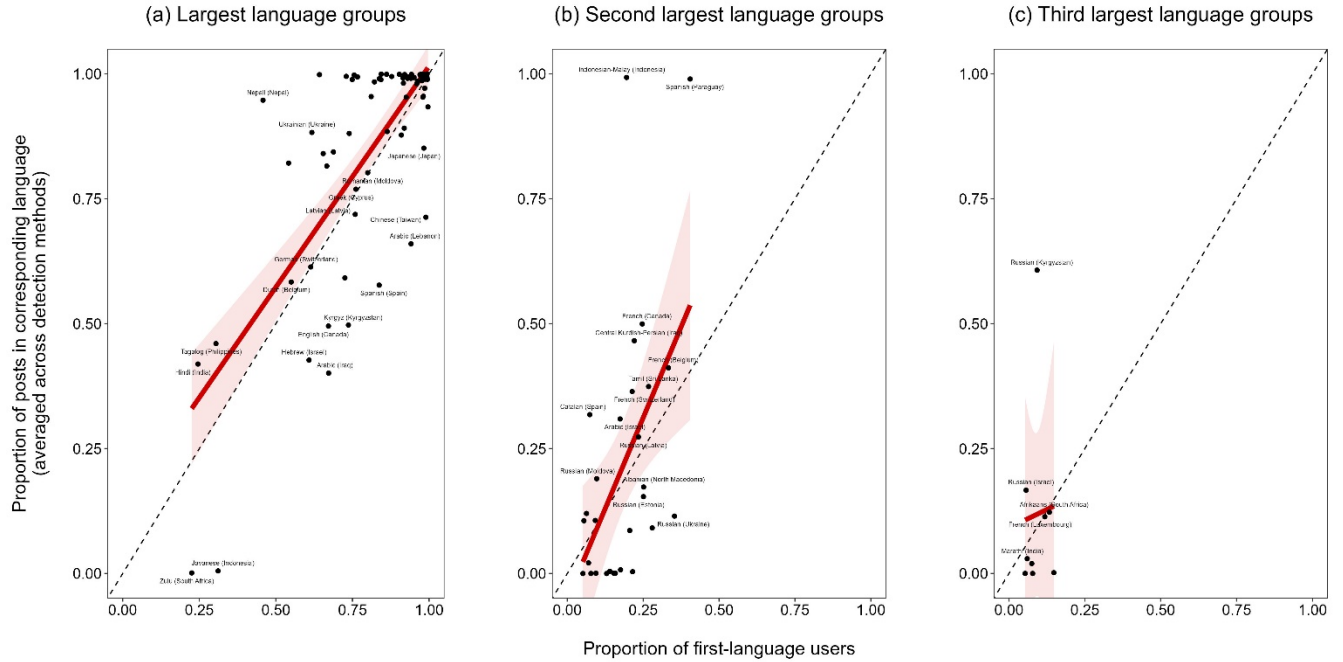
Note: †p<0.1; \*p<0.05; \*\*p<0.01. The models are estimated on a linear probability model with standard errors clustered by country. The observations are weighted by the logged number of posts. ELF = ethno-linguistic fractionalization.

Table 3: Primarily multilingual parties and the language use of candidates

	Primarily multilingual candidate		Minority language candidate	
	(1)	(2)	(3)	(4)
Primarily multilingual party	0.24** (0.04)	0.22** (0.05)	0.19** (0.06)	0.13 <sup>†</sup> (0.08)
Social ideology		0.01 (0.01)		-0.01 (0.01)
Majoritarian system		-0.09 (0.07)		0.07 (0.09)
ELF		0.19 <sup>†</sup> (0.11)		0.02 (0.14)
$\widehat{\sigma}_p$	0.14	0.13	0.25	0.22
$\widehat{\sigma}_c$	0.12	0.05	0.03	0.00
N of candidates	4,188	3,452	4,188	3,452
N of parties	111	67	111	67
N of countries	12	12	12	12

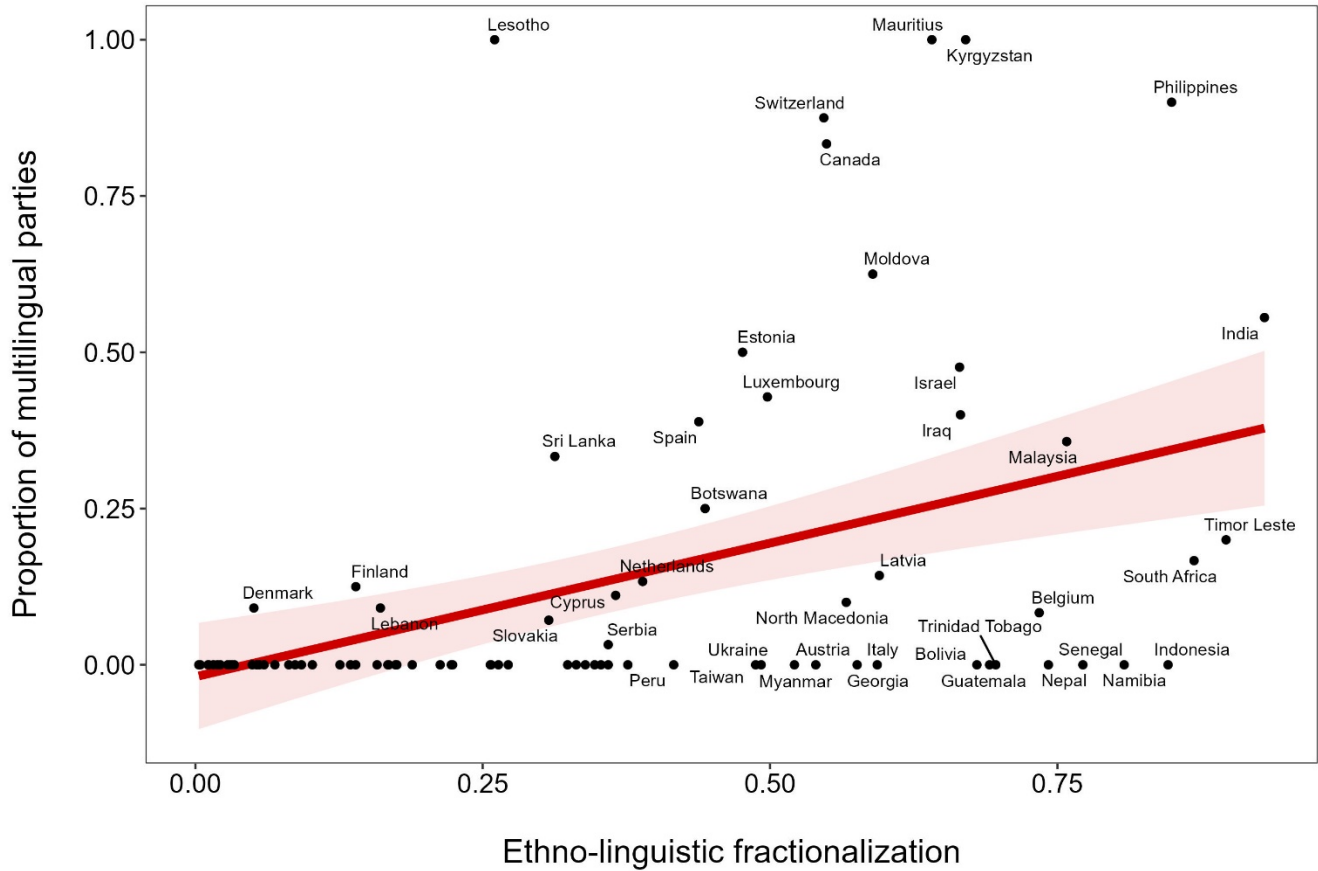
*Note:* <sup>†</sup>p<0.1; \*p<0.05; \*\*p<0.01. The models are estimated on a linear model with nested random effects by party and country. Observations are weighted by the logged number of posts.  $\widehat{\sigma}_p$  and  $\widehat{\sigma}_c$  indicate the estimated standard deviation of party and country random effects. ELF = ethno-linguistic fractionalization.

Figure 1: Relationship between the proportion of first-language users and the proportion of posts in corresponding languages



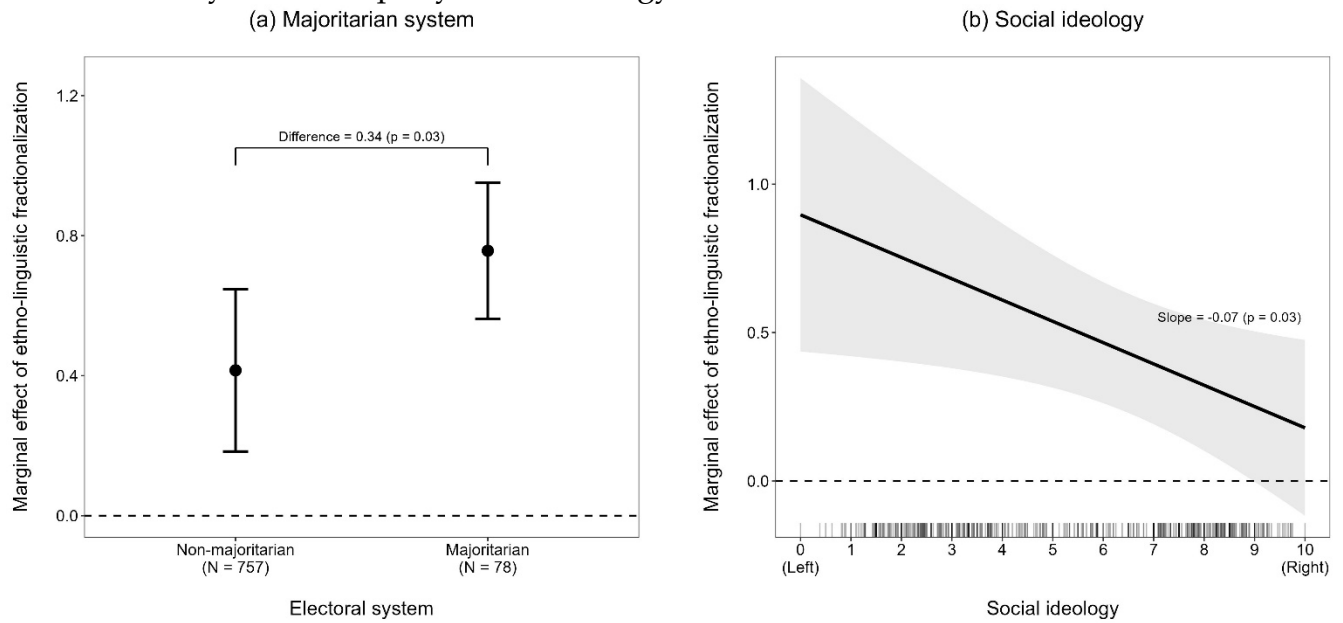
*Note:* The figure shows the relationship between the proportion of first-language users and the proportion of party Facebook posts written in the corresponding language. Dashed lines are 45-degree lines, and solid lines are estimated on a bivariate linear regression. Shaded areas indicate 95% confidence intervals. The figure only reports languages detected in three or more detection methods.

Figure 2: Linguistic diversity and the proportion of primarily multilingual parties



*Note:* The figure shows the relationship between linguistic diversity and the proportion of multilingual parties by country. The solid line is estimated on a bivariate linear regression. Shaded area indicates a 95% confidence interval.

Figure 3: Marginal effects of linguistic diversity on parties' multilingualism conditional on electoral systems and party social ideology



*Note:* Panel (a) summarizes the marginal effect of linguistic diversity on parties' multilingualism conditional on electoral systems based on Model 1, Table 2. Panel (b) shows the marginal effect of linguistic diversity on parties' multilingualism conditional on social ideology based on Model 2, Table 2. Vertical bars and shaded areas indicate 95% confidence intervals.