Z Vgl Polit Wiss (2020) 14:425–459
https://doi.org/10.1007/s12286-020-00472-3

ZEITSCHRIFT FÜR VERGLEICHENDE POLITIKWISSENSCHAFT

AUFSÄTZE

Democracy in context: using a distributional semantic model to study differences in the usage of democracy across languages and countries

Stefan Dahlberg · Sofia Axelsson · Sören Holmberg

Accepted: 4 December 2020 / Published online: 18 December 2020

Abstract Cross-cultural survey research rests upon the assumption that if survey features are kept constant, data will remain comparable across languages, cultures and countries. Yet translating concepts across languages, cultures and political contexts is complicated by linguistic, cultural, normative or institutional discrepancies. Such discrepancies are particularly relevant for complex political concepts such as democracy, where the literature on political support has revealed significant cross-cultural differences in people's attitudes toward democracy. Recognizing that language, culture and other socio-political variables affect survey results has often been equated with giving up on comparative research and many survey researchers have consequently chosen to simply ignore the issue of comparability and measurement equivalence across languages, cultures and countries. This paper contributes to the

Text data at word level and certain metadata, i.e. data output from the models, that support the findings of this paper are available from the corresponding author upon reasonable request. Raw text data at web document level, i.e. input data used to build the models, are not publicly available due to restrictions from a third-party data provider. All models were built in Python using the standard setting of the Gensim library, an open-source software available via https://pypi.org/project/gensim/ [accessed 20 November 2020] under copyright of Radim Řehůřek. Data management and analysis of text data output from the models was implemented using R and Stata, and code that support the findings of this paper is available from the corresponding author at reasonable request.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12286-020-00472-3) contains supplementary material, which is available to authorized users.

Prof. S. Dahlberg
Department of Humanities and Social Sciences, Mid Sweden University, Östersund, Sweden
E-Mail: stefan.dahlberg@miun.se

S. Axelsson (✉) · Prof. S. Holmberg
Department of Political Science, University of Gothenburg, Gothenburg, Sweden
E-Mail: sofia.axelsson@gu.se

Prof. S. Holmberg
E-Mail: soren.holmberg@pol.gu.se

debate, using a distributional semantic lexicon, which is a statistical model measuring co-occurrence statistics in large text data. The method is motivated by structuralist meaning theory, stating that words with similar meanings tend to occur in similar contexts, and that contexts shape and define the meanings of words. Compared to other methodological approaches aimed at identifying and measuring cross-cultural discrepancies, this approach enables us to systematically analyze how the concept of democracy is used in its natural habitat. Collecting geo-tagged language data from news and social online source documents this paper descriptively explores varieties in meanings of democracy across a substantial number of languages and countries, and maps ways in which democracy is used among online populations and regions worldwide.

Keywords Cross-cultural surveys · Distributional semantics · Meaning of democracy · Online text data · Translation discrepancies · Word embeddings

Demokratie im Kontext: Verwendung eines verteilungssemantischen Modells zur Untersuchung von Unterschieden in der Verwendung des Wortes „Demokratie“ über Sprachen und Länder hinweg

Zusammenfassung Die interkulturelle Umfrageforschung beruht auf der Annahme, dass bei konstanten Umfragemerkmalen die Daten über Sprachen, Kulturen und Länder hinweg vergleichbar bleiben. Die Übersetzung von Konzepten über Sprachen, Kulturen und politische Kontexte hinweg wird jedoch durch sprachliche, kulturelle, normative oder institutionelle Diskrepanzen erschwert. Besonders für komplexe politische Konzepte wie Demokratie, sind solche Diskrepanzen jedoch höchst relevant, da die Literatur zur politischen Unterstützung erhebliche interkulturelle Unterschiede in der Einstellung der Menschen zur Demokratie bereits offenbart hat. Die Erkenntnis, dass Sprache, Kultur und andere sozio-politische Variablen die Ergebnisse von Umfragen beeinflussen, hat entweder dazu geführt, dass vergleichende Survey-Forschung nicht mehr durchgeführt wurde, oder dass viele Umfrageforscher sich entschieden haben, die Frage der Vergleichbarkeit und der Äquivalenz der Messungen über Sprachen, Kulturen und Länder hinweg einfach zu ignorieren. Dieses Papier leistet einen Beitrag zu dieser Debatte, indem es ein verteilungsbezogenes semantisches Lexikon verwendet, bei dem es sich um ein statistisches Modell zur Messung von Koinzidenzstatistiken in großen Textdaten handelt. Die Methode ist durch die strukturelle Semantik motiviert, die besagt, dass Wörter mit ähnlichen Bedeutungen tendenziell in ähnlichen Kontexten vorkommen und dass Kontexte die Bedeutungen von Wörtern formen und definieren. Im Vergleich zu anderen methodischen Ansätzen, die darauf abzielen, kulturübergreifende Diskrepanzen zu identifizieren und zu messen, ermöglicht uns dieser Ansatz eine systematische Analyse, wie das Konzept der Demokratie in seinem natürlichen Lebensraum verwendet wird. Das Papier sammelt geo-markierte Sprachdaten aus Nachrichten und sozialen Online-Quelldokumenten, untersucht deskriptiv die Vielfalt der Bedeutungen von Demokratie in einer beträchtlichen Anzahl von Sprachen und Ländern, und kartographiert die Art und Weise, wie Demokratie in Online-Bevölkerungen und Regionen weltweit verwendet wird.

Schlüsselwörter Interkulturelle Umfragen · Verteilungssemantik · Demokratievorstellungen · Online-Textdaten · Übersetzungsdiskrepanzen · Worteinbettungen

1 Introduction

In medieval art, Moses is often depicted as a man with horns instead of a man with divine radiance. The legend tells that this is the result of a translational mistake in Exodus, chapter 34, where Moses returned from Mount Sinai after receiving the Ten Commandments. It is said that it was Saint Hieronymus who mistakenly translated the Hebrew word *qaran* into the Latin word *cornuta* ("horned") instead of the correct term *coronata* ("radiant"). This is perhaps one of the most famous translational oversights in history. Today, we do not need to rely (solely) on the Bible to inform us about human nature. Instead, we can consult data from any of the large-scale country comparative survey programs including but not limited to the International Social Survey Programme (ISSP 1984–), the World Values Survey (WVS 1990–), the Comparative Study of Electoral Systems (CSES 1996–) and the Global Barometers (2004–). Together, such surveys cover a majority of the world's population on various themes of relevance for increasing our understanding of the world. Still, not even surveys are exempt from translational oversights, which can lead to discrepancies in how people perceive a specific survey item or survey question (Hoffmeyer-Zlotnok and Harkness 2005).

Gauging attitudes toward multifaceted concepts in a cross-cultural context is complicated as its implementation requires measurement equivalence across languages, cultures and countries (King et al. 2004). Language—an instant and inevitable representation of culture—activates cognitive frames that are linked to cultural understandings of concepts, and in a survey setting, what is generally referred to as language effects occur when the language of administration affects the ways in which respondents answer survey questions (Zavala-Rojas and Saris 2018). Although it is sometimes possible to unambiguously translate lexical items across languages—and even as survey methodologist are paying increasing attention to the role of translation—concepts that pertain to complex socio-political phenomena can still evoke different meanings and associations regardless of the survey translation (Behling and Law 2000; Braun and Harkness 2005).

One of the most complex concepts of contemporary political science is no doubt that of democracy. Originating from the Greek *dēmokratia*, democracy was established via late Latin in modern vocabulary in the 16th century, and many languages have adopted versions of the original word that have long been used in survey questions about democracy. Although democracy literally means rule by the people—coined from *dēmos* ("people") and *kratos* ("rule"), the word and its many translations have been shown to carry different meanings depending on the cultural context, and such meaning discrepancies are not easily identified by comparative survey methodology (Schaffer 2000; cf. Welzel 2013).

Recognizing that language and culture affect survey results has often been equated with "giving up on comparative research" and, consequently, the most commonly

"solution" to equivalence problems has been for researchers to simply ignore the issue of measurement comparability across languages, cultures and countries (King et al. 2004). This paper contributes to the debate by using distributional semantics to account for semantic differences between lexical realizations of concepts across languages. Distributional semantics is a statistical approach for quantifying semantic similarities based on co-occurrence information collected from large text data (Turney and Pantel 2010). In this experiment, we explore how the word democracy is used in geo-coded language data from online news and social media sources. The reason for using such data rather than balanced corpora is that it enables us to analyze word meanings in normal, uncontrolled, unsolicited, and contemporary language use. Compared to other methodological approaches aimed at identifying and measuring cross-cultural discrepancies, this approach has the advantage of enabling us to analyze how concepts are used in their "natural habitat" (Wittgenstein 1958). The focus of this paper is not primarily that of correcting the various translations of democracy that are used in surveys but to explore how the word democracy and its translations are used in online text data within and between different cultural regions. Our main aim is to present a novel approach to measuring the meaning of democracy and, doing so, our ambition is to uncover potential meaning differences across languages and countries that may be beneficial for the survey-based community of democracy research in the long run.

2 Anchoring the "D-word" in a cross-cultural setting

For years, survey programs have asked citizens questions about democracy; whether they support democracy as an ideal form of government, how democratic their country is, or how satisfied they are with the ways in which democracy is working is their country. Citizens' attitudes toward democracy remains a major topic in political science, and recent democratic setbacks in several diverse countries including Brazil, Hungary, Russia and Turkey as well as electoral success of populist leaders such as Donald Trump in the United States and Rodrigo Duterte in the Philippines and the rise of the populist radical right in Europe, have brought the issue of democratic consolidation to the fore even more (e.g. Dalton 2004; Foa and Mounk 2016, 2017; Pharr and Putnam 2000; Mounk 2018; Norris 2017).

It has long been assumed that the legitimacy of any democratic regime is contingent on the public's support for democracy (Diamond 1999; Lipset 1959). The importance of attitudes and assessments for democratic legitimacy is a cornerstone in the substantial body of literature devoted to measuring public opinion in relation to democracy (Inglehart and Welzel 2005; Klingemann 1999). The concept of political support, originally conceptualized by Easton (1975) generally differentiates between "diffuse support" for the political community and for democratic principles on the one hand, and "specific support" for the specific democratic regime and the outcomes delivered by its institutions and actors (Norris 1999). According to this conceptualization, citizens can certainly be critical of the incumbent democratic regime or dissatisfied with government performance whilst simultaneously support-

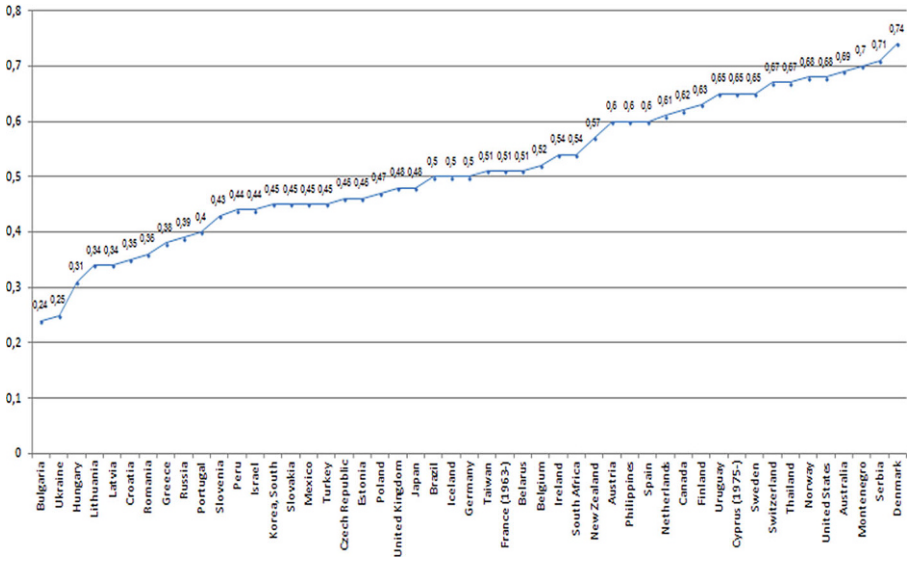


Fig. 1 Satisfaction with the Way Democracy Works Across 49 Countries (The aggregated measures of citizen's satisfaction with the way democracy works (SWoD) are based on data from two different data sources. The Comparative Studies of Electoral Systems (CSES) Modules 3 and 4 (2006–2016) and the European Social Survey (ESS) Wave 3 (2008). In both surveys, the question reads: "On the whole, how satisfied are you with the way democracy works in [country]?" In contrast to the CSES questionnaire (where the response options are 1—not at all satisfied to 4—very satisfied), the ESS response options are based on an 11-point scale, ranging from 0 (extremely dissatisfied) to 10 (extremely satisfied) (for more information, see www.europeansocialsurvey.org/data). Differences in scale and time are not optimal for comparisons. However, for 23 countries, data were overlapping between CSES and EES and the correlation between the two survey measures was $r = 0.81$, which makes them not identical but at least very close. Based on this correlation we have combined them into one dataset where country averages were rescaled into 0–1 with high values indication satisfaction.)

ing democracy as the ideal form of government and adhere to democratic principles and values (Inglehart 2003).

Yet, there remains much conceptual confusion about political support resulting in direct and indirect indicators becoming conflated and used indistinctly to capture various dimensions of democracy (Kriesi et al. 2013). One example is the satisfaction with democracy (SWoD) indicator, which have frequently been used as a proxy for support for democracy (Canache et al. 2001). A study by Linde and Ekman (2003) for instance reveals that the SWoD indicator captures regime performance—particularly subjective economic measures—rather than regime legitimacy. Additional research has supported these findings and carefully concluded that performance factors including institutional quality and quality of government services appears to be strong determinants of satisfaction, which suggests that the SWoD indicator reflects popular support for how the democratic regime works in practice (Dahlberg and Holmberg 2014).

Fig. 1 shows the aggregated levels of citizens' satisfaction with the way democracy works across 49 countries, and is based on data from two different survey sources, the Comparative Studies of Electoral Systems (CSES) Modules 3 and 4

(2006–2016) and the European Social Survey (ESS) Wave 3 (2008). Surprisingly, both Serbia and Montenegro rank second and third in terms of highest satisfaction with democracy, despite being significantly flawed democracies (Lührmann et al. 2019). We also observe that Thailand—a closed autocracy—precedes both Switzerland and Sweden, which are generally ranked among the top 10% most liberal democratic countries in the world (ibid.). Other scholars have observed similar variations as in Fig. 1 for other survey questions measuring self-assessments of democracy where citizens in illiberal regimes like Azerbaijan, Belarus, China, Russia or Vietnam either believe their country to be democratically governed to a substantial degree (Ariely 2014) or report widespread support for democracy (Welzel and Klingemann 2008; Welzel and Kirsch 2017).

Not only do such findings raise concern about the validity and comparability of indicators related to measure attitudes and assessments of democracy, but they make us question whether there is in fact a universal understanding of democracy, or if citizens in fact attribute different meanings to the concepts (Cutler et al. 2013). When the word democracy is used cross-culturally, what does it actually mean? In the following sections, we discuss some of the most important findings on the meaning of democracy from recent quantitative as well as qualitative work and present a methodology that leverages on both approaches to anchor the "D-word".

2.1 The systematic survey approach

People all over the world are widely supportive of democracy (Norris 1999), but its conceptual vagueness allows for different understandings and usages of the concept (Schedler and Sarsfield 2007), and its "polysemic nature" is reflected in the comparative political culture literature (Ariely 2014, p. 624).

Studies have found that democracy indicators tend to be more performance-oriented in new democracies, particularly former Soviet states, compared to established democracies (Holmberg 2014; Dahlberg et al. 2015). Mishler and Rose (2001) distinguish between an idealist approach and a realistic approach in measuring support for democracy, arguing that citizens in newly democratized countries may have limited familiarity with democratic ideals, and values and therefore a lessened capacity to assess democracy in abstract terms. Although they may certainly develop the capacity for more abstract assessments of democracy over time, contingent on the progress delivered by the new regime, their real-lived experience makes them better equipped to evaluate immediate regime performance. Their study on Post-Communist states show that realist measures of democratic support are preferable to idealistic ones.

Bratton and Mattes (2001) reach somewhat different conclusions when they explore whether support for democracy is intrinsic—the embodiment of democratic values and of democracy as objective in itself—rather than instrumental—economic gains and improved material living standards. Examining the intrinsic versus the instrumental debate using survey data from Ghana, Zambia and South Africa, the authors find that both notions matter; respondents were generally dissatisfied with government achievements but nonetheless expressed attachment to democratic norms and values. Although approval of the democratic regime was found to be contingent

on government performance, economic performance proved to matter less for citizens compared to political performance including ensuring political rights and civil liberties.

Other scholars argue that a universal understanding of democracy may in fact exist, or at least that people "define democracy on the basis of common criteria" (Ariely 2014, p. 624). Using open-ended survey question asking people to themselves define the meaning of democracy, Dalton et al. (2007) find that most respondents—also those living in developing democracies—are more than capable of ascribing meaning to the concept of democracy, with a majority defining it in terms of political freedoms, civil liberties and rights rather than in terms of institutional or procedural features, and socio-economic features including development, peace and security (see also Shin 2017).

In recent years scholars have made significant efforts to measure "varieties of democracy" by conceptualizing and operationalizing key principles that represent different approaches to democracy (e.g. electoral, liberal, participatory, deliberative and egalitarian) in order to create comprehensive cross-national democracy indices over time (Lindberg et al. 2014). Despite a thorough methodology, such indices nonetheless rely on the assessments of country experts, and not on public perceptions. Moreover, as Kriesi et al. point out, they "measure the quality of existing democracies against [a] ... theoretical yardstick of what democracy ought to be", not what it actually means to people across the world (2013, p. 1).

The World Values Surveys (WVS) have tried to remedy the issue of potential bias induced by scholars that already adhere to a liberal democratic definition of democracy by including a survey battery with questions that pertain to four distinctive notions of democracy: a liberal, a social, a populist, and an authoritarian notion. In several studies, Welzel leverages on the WVS questions in order to explain "the paradox of democracy"—that overwhelming support for democracy paradoxically tends to coexist with a lack of democracy altogether (Welzel 2013; Welzel and Kirsch 2017). Examining support for democracy under the control of the distinct notions, the paradox disappears; in places where democracy is deficient or simply absent, citizens' perceptions of what democracy means is distorted in favor of authoritarianism. This is particularly evident for countries where citizens lack the emancipative values required in order to go from desire for democracy to concrete action for democracy (2013, p. 330). Consequently, Welzel and Kirsch explain, "authoritarian *mis*understandings of democracy might be widespread and real ... under false notions of democracy, people consider non-democratic regime characteristics as democratic" (2017, p. 3).

From this, we acknowledge that democracy appears an "essentially contested concept" (O'Donnell 2007; in Ariely 2014, p. 624), and that the systematic survey approach, despite its vast and nuanced empirical literature, is not always equipped to assess the scope of cross-cultural variations in the meaning of democracy in different cultural settings.

2.2 The exploratory ethnographic approach

Turning to the work in the domain of qualitative methodology, Schaffer (2000) offers one of the most comprehensive efforts to anchor the "D-word" in a cross-cultural setting. Whereas survey-based scholarship on perceptions of democracy frequently focus on institutional factors such as regime type and regime age, Schaffer takes on a semantic approach to investigate the issue of cross-cultural comparability in relation to democracy. He argues that cross-cultural analyses of attitudes toward democracy must take into account both (1) whether the institutions of the countries in question are comparable, and (2) the ideals, values and standards attributed to those institutions. This allows for an analysis of how the meaning people ascribe to democratic institutions may vary across contexts. More specifically, Schaffer suggests that the meaning of democracy can be traced through language by using conceptual analysis, which considers the structure of the concept, its associated meanings, its use in everyday language, and how the concept fits into a "semantic field" of related concepts. In doing so, Schaffer emphasizes that the meaning of a concept is best captured by carefully studying how it is used in its everyday context.

Schaffer applies this method across languages in order to identify similarities and differences in the meaning attributed to democracy in two seemingly very different countries; the United States and Senegal. Both countries, however, have a long tradition of competitive elections, which makes it possible to assume that the Senegalese have roughly similar ideas of democratic institutions, although the countries differ in terms of social organization, cultural and religious traditions as well as political practices. With this case selection, Schaffer controls for Dahl's (1989) proposition that new and old democracies differ in their conception of democracy. He compares the semantic fields of the French word *démocratie*, used by the French-speaking population in Senegal, and the Wolof word *demokaraasi*, used by the largest ethnic group, against the American English word *democracy* by tracing usages of the concepts in the political arena during democratic elections in Senegal and the United States, more specifically so by conducting open-ended interviews, attending party congresses and political rallies, and examining television and radio broadcasts.

Schaffer finds that French concept largely mirrors the English where democracy is associated with distributive equality, inclusive participation and meaningful choice. The Wolof concept, however, is related to concerns about collective economic security and community loyalties. In fact, whilst *demokaraasi* as concept does have things in common with its French and English counterparts—welfare and electoral participation—*demokaraasi* as a practice generally refers to solidarity, consensus and even-handedness, which many times override the typical Western ideals.

Schaffer subsequently compares the Wolof understanding of democracy with results from similar studies conducted in other parts of the world. He cites earlier work on the Chinese concept *minzhu* (a common translation of democracy) which implies popular participation under elite supervision, promotion of the common interest, and public scrutiny of the workings of bureaucracy—elements that are all supposed to work in favor of national unity. Regardless of its seemingly authoritarian notion, *minzhu* was nonetheless a term used in the student-led Tiananmen Square protests

of 1989, demanding democratic reform. Although the Chinese, Wolof and English concepts of democracy carry different meanings, these meanings also partially overlap: *minzhu* and *democracy* share a notion of popular political participation, while *demokaraasi* shares with *minzhu* the notion of unity. Following Wittgenstein, Schaffer suggests that we could conceive of these conceptual relations as family resemblances; "as the pattern of overlapping and crisscrossing similarities (...) between the ways in which roughly equivalent words get used in different languages" (2000, p. 145).

Importantly, by exploring the use of language, Schaffer shows that citizens in different cultures speaking different languages reflect different understandings of democracy that may not be beneficial for the stability of democracy in the long run. For instance, even as many Wolof-speakers are engaged in democratic practices and desire electoral participation, their endorsement of consensus politics and collective solidarity, coupled with economic uncertainty and clientelism, may not result in electoral accountability after all. Instead of individual participation and inclusivity of the majority population, Senegalese election results may in part reflect the opposite (Kasfir 2000).

2.3 The nexus approach

The different approaches have their advantages, but also limitations and caveats; survey-based studies allow for global statistically supported comparisons, but many existing survey items suffer from validity issues as it has proven difficult to establish whether democracy means the same to people across linguistically, culturally and socio-politically diverse societies. Ethnographic studies in contrast allow for "thick description" and enhance our understanding of what democracy means for people in ordinary socio-political life. This approach captures both political and non-political uses of democracy, which can be used as an indicator of the degree to which the concept is anchored in society. However, it is by default limited in its scope, which undermines the generalizability of findings in a comparative setting.

This paper combines the systematic approach of comparative surveys with the exploratory approach of ethnographic studies, and while we cannot offer any definite solutions to the issues of measurement comparability and cross-cultural generalization inherent in respective approaches, we do believe that this nexus approach represents an innovative way forward as it makes it possible to identify how democracy across cultures without using any priming techniques.

Utilizing recent advances from distributional semantics, we focus on the meaning of democracy in online text data, exploring which other words are semantically similar to the word democracy and whether the usages of the word democracy differs across countries and cultural regions. Because language is considered a proxy for culture in that it triggers culturally-specific understandings of concepts, and has proven to evoke different meanings of democracy, we examine common translations of democracy in their respective languages. Existing geo-coding enables us to discern the country origin of the language data, which is based on texts from both online news and social media sources.

Before we turn to the methodology section of this paper, it should be added that whilst the method by which we collect data and measure semantic similarity is systematic, the subsequent analysis is largely explorative. The main aim of this paper is to present an alternative approach to measuring the meanings of democracy, and our findings are purely descriptive at this point. Even so, we still find it reasonable that a novel methodology for identifying meaning differences may prove beneficial to the survey-based community. Not only do we believe that a methodology of this kind can be used to estimate the effectiveness of different survey translations, but its findings can also be used to develop and improve survey design and question content.

3 Distributional semantics as method

In order to study how the word democracy and its many translations are used, we have used distributional semantics to build an online lexicon that is based on text data from geo-coded news and social web documents (for more detailed description of the lexicon and the methodology behind the lexicon, see Dahlberg et al. 2020). Distributional semantics is an area of natural language processing, centered on the study of various forms of natural language modeling. It is grounded structural meaning theory and often summarized in the words of one of its founding fathers, John Rupert Firth: "You shall know a word by the company it keeps" (1957, p. 11). Studying the meaning of a word requires us to "specify under which conditions two words can be said to have the same meaning or—if we regard the notion of synonymity too strong—to be semantically similar" (Lenci 2008, p. 2). According to Lenci (2008), the theoretical assumption of any distributional semantic model is the definition of semantic similarity as linguistic distributions. This has become widely recognized as the Distributional Hypothesis, popularized by Firth (1957), which Lenci formulates in the following way: "The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B appear" (2008, p. 3). Or, as Sahlgren puts it: "if we observe two words that constantly occur in the same contexts, we are justified in assuming that they mean similar things" (2005, p. 1).

Distributional semantic models collect co-occurrence statistics from large dynamic text data (often referred to as Big Data) in order to produce a multidimensional vector space in which each word is assigned a corresponding vector (Sahlgren 2006, 2008). Word vectors are positioned in the vector-space so that words that share a common context are located in close proximity to one another in the vector-space. Relative distance between word vectors, measured as cosine similarity ranging from –1 to 1, indicates the degree of similarity of usage between words. In this way, distributional semantic models can be used to find semantically similar terms to a given target or query term and, in effect, a distributional semantic model constitutes a statistically compiled lexicon.

For instance, a distributional semantic model would likely return words like "green", "yellow", "black", and "white" when probed with the word "red". Applying such models to different text sources can tell us which words are used in

similar ways and effectively give us a clear indication of unsolicited word usages that we are unable to find in standard lexical or conceptual resources. Computing and comparing such relations across languages will tell us about similarities and differences in the meaning of words, and will highlight translation discrepancies. Recalling Wittgenstein's notion of "meaning as use"—emphasizing that formal relations between linguistic items are meaningless outside the context in which they are used—the method allows us to investigate how the word democracy is used in its "natural habitat" (1958).

3.1 Data and modeling

Though natural language—generally defined as language that has naturally developed over time without any premeditation (Lyons 1991)—can take on a variety of forms including spoken language and signed language, the data analyzed herein is that of written language, more specifically, online text data. The data is derived from a commercial partner, the Swedish language technology company *Gavagai*, which utilizes a number of large-scale commercial data providers that regularly crawl data from numerous freely available internet domains across the world.

The commercial partner provides us with vast batches of structured online text data and metadata in the form of indexed web documents—e.g. a blog post, a forum post, or a news article—that are classified by source language, source country and source type. The source URL has already been classified by the crawlers, and is defined as either news data, forum data, or social data. As it is sometimes difficult to distinguish forum data from social data, we have combined the two and simply refer to this combination as social data.

In terms of distributional semantic models, there exist many with different properties that can be used to derive word vectors. The choice of model depends on the nature of the inquiry—the type of similarity relations one is seeking to compute as well as the size of data or frequency range of terms—and is generally a question of performance versus efficiency (Sahlgren and Lenci 2016). Because of desirable properties, such as high-quality vectors with respect to semantic similarity and efficient computational implementation, we have selected the Continuous Bag of Word (CBOW), a word2vec model introduced and developed by Mikolov et al. (2013).

For the CBOW model, a number of model and training parameters that determine its performance including window size, embedding dimension and vocabulary size. Since we are interested in paradigmatic relations where words in the same category can be substituted with one another, we choose parameters that reflect this notion of semantic similarity (Sahlgren 2006). For this reason, the window size has been set to three as it has been shown to be a suitable size for modeling paradigmatic relations (Levy and Goldberg 2014; Karlgren and Sahlgren 2001). For the embedding dimension, we have used a size of one hundred, which is a compromise between model size and model quality. Given that we train and store a large number of models—one for each language, country and source type—the model size directly impacts on the responsiveness of the server that is hosting all models, and by extension also its interface—i.e. the online lexicon—used for collecting output data from the models. Therefore, we have chosen a smaller, albeit frequently used, embedding dimension.

With regard to the vocabulary size, the models have been trained on at least 1 GB of text data, where words that occur less than five times have been dropped from the vocabulary as per standard practice. Because the amount of text data for every language, country and source type differs quite a lot in size, and the total amount of data available is not necessarily known, each model randomly samples text, we start sampling a small proportion and double it until we reach 1 GB of text data. For models where the data is scarce but nonetheless more than 1 GB, we sample all available data (see Dahlberg et al. 2020).[1]

Before training the models, the sampled texts go through some additional preprocessing including tokenization and lowercasing. We have applied the same tokenizer for most languages, namely, the NLTK casual tokenizer tool[2]. For a few languages, we have used separate tokenization tools: Stanford NLP Word Segmenter[3] for Arabic and Chinese (using different classifiers), PyThaiNLP[4] for Thai, and the Python Vietnamese Toolkit[5] for Vietnamese. All models are implemented in Python using standard settings of the Gensim library by Řehůřek and Sojka (2010).

It is well documented that many distributional semantic models (or word embedding models as they are often referred to) are not stable; the same model, with the same parameters, trained on the same data, will yield different results. As Antoniak and Mimno (2018) have shown, models differ in their degree of variability—it is not always clear whether they reflect continuous properties of language use rather than source document specificities—especially when trained on slightly different data sets. However, it should be noted that the median corpus size for all our source-language-country corpora (~140,000,000 words for all news data and ~110,000,000 words for all social data) is an order of magnitude larger than the largest corpus investigated by Antoniak and Mimno and that larger corpus size significantly lower this kind of word embedding variance (Pierrejean and Ludovic 2018).

While we do not bootstrap the models, the analysis presented in the findings section is performed at a highly aggregated level both at model level, corpora level (where we aggregate national estimates to regional estimates) and word level (where we aggregate words into topics), which effectively lowers the effect of individual model variance. Stability is indeed a desirable property but it is trumped by word embedding quality—after all, it is trivial to construct stable yet semantically nonsensical word embeddings (Rogers et al. 2018). For our purposes, we have chosen CBOW as a compromise between performance and computational expedience. Although Antoniak and Mimno (2018) find that the GloVe model (Pennington et al. 2014) is more stable than other models, models such as CBOW and SkipGram (Mikolov et al. 2013) are better able to model paradigmatic relations, low frequency

[1] More details on the methodology behind the lexicon including data and modeling is available in the unpublished report *The LES Distributional Semantic Lexicon* by Dahlberg, Stefan, Sofia Axelsson, Magnus Sahlgren, Amaru Cuba Gyllensten, Ariel Ekgren and Sören Holmberg (2020), which can be accessed upon request from the corresponding author.

[2] See https://www.nltk.org/api/nltk.tokenize.html.

[3] See https://nlp.stanford.edu/software/segmenter.html.

[4] See https://github.com/PyThaiNLP/pythainlp.

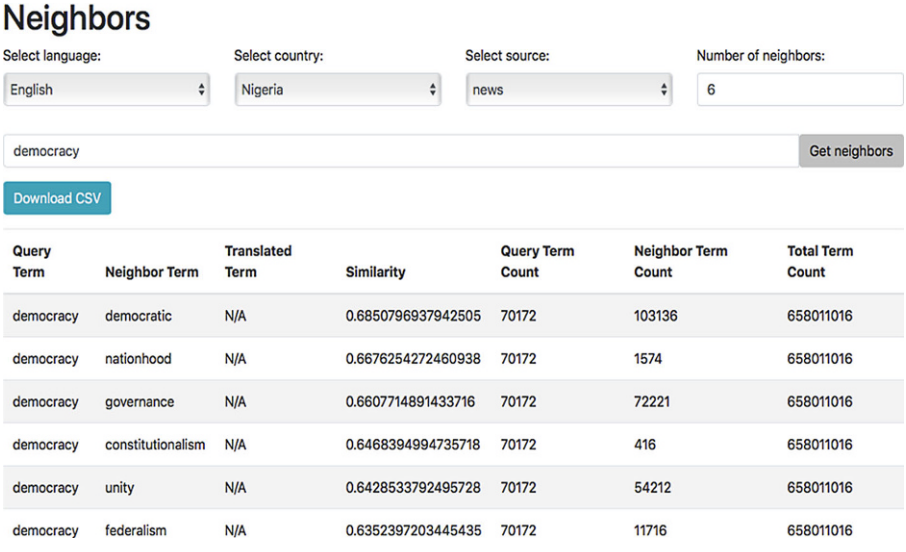
[5] See https://github.com/trungtv/pyvi.

neighbors as well as close neighbors, and are generally better than GloVe when it comes to downstream tasks. Moreover, while the SkipGram model appears to perform better than the CBOW model, the latter has proven more computationally efficient than the former (Rogers et al. 2018), an important factor considering that we train a total of 228 different embedding models (see Dahlberg et al. 2020).

3.2 Collecting and translating online text data

Together, these models make up a vast online distributional semantic lexicon covering approximately 140 language-country combinations. To simplify the data collection, we have built an interface, which effectively works as an online lexicon where we can query any given model and export the data returned from the query (Dahlberg et al. 2020). Fig. 2 shows what the lexicon looks like in practice after we have asked it to compile the six most semantically similar terms—in computational linguistics also referred to as neighbor terms—to the term democracy in English news data from Nigeria.

In December 2017, we exported data 15 neighbor terms from news data and 15 from social data for 114 language-country combinations from the lexicon using the query word democracy in respective languages, and all terms were subsequently translated. We make use of the phrase language-country combinations rather than



Query Term	Neighbor Term	Translated Term	Similarity	Query Term Count	Neighbor Term Count	Total Term Count
democracy	democratic	N/A	0.6850796937942505	70172	103136	658011016
democracy	nationhood	N/A	0.6676254272460938	70172	1574	658011016
democracy	governance	N/A	0.6607714891433716	70172	72221	658011016
democracy	constitutionalism	N/A	0.6468394994735718	70172	416	658011016
democracy	unity	N/A	0.6428533792495728	70172	54212	658011016
democracy	federalism	N/A	0.6352397203445435	70172	11716	658011016

Fig. 2 Online Distributional Semantic Lexicon (To clarify the columns of the matrix, it should be noted that the column 'Query Term' returns the term used to query the lexicon for semantically similar terms. 'Neighbor Term' returns the semantically similar terms in rank order, based on the similarity score presented in the column 'Similarity'. The column 'Translated Term' is connected to the API of Google Translate and provides translations for other languages than English. 'Query Term Count' returns the frequency of the query term in the modeled corpus (in this case, English news data from Nigeria), and 'Neighbor Term Count' returns the frequency of each unique neighbor term in the modeled corpus. The column 'Total Term Count' is the total amount of terms available in the modeled corpus. Please note that this example was generated from the lexicon in 2019.)

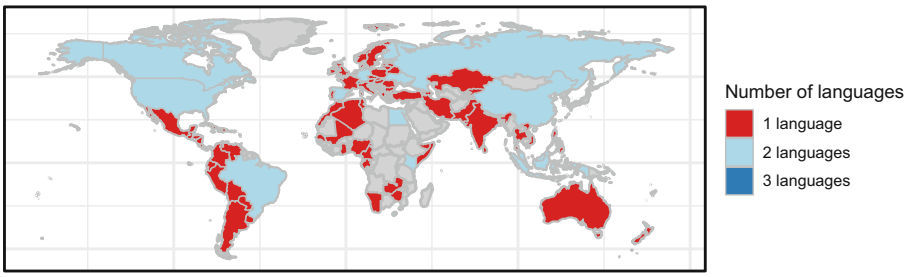


Fig. 3 Online Text Data Coverage (Countries marked "1 language" represent those from which we have extracted text data in one language (e.g. Chile with Spanish only), countries marked "2 languages" represent those from which we have extracted text data in two languages (e.g. Canada with English and Spanish), and the country marked "3 languages" is the one from which we have extracted text data in three languages (Switzerland with French, German and Italian).)

countries as we have collected data in more than one language for several countries. For instance, we have collected neighbor terms in French, German and Italian from Switzerland, and English and French from Canada. After the translation process, 3420 original neighbor terms were reduced to 3345 were a few language-country combinations from either news or social source types were omitted due to poor data quality. The data coverage is visualized in Fig. 3, and a complete overview of all target terms and translated neighbor terms for all languages and countries are available in Table 6 (online) in the appendix.

Knowing that machine translators cannot guarantee the interpretive sophistication required for this type of study, a large group of international research assistants were employed to assist with the translation process. In view of the language agnostic approach of this paper, this is somewhat methodologically fragile given that human translators inevitably introduce bias to the material. However, the translations were conducted in a supervised environment, where the translators were instructed not only to provide translation suggestions to the terms but also to thoroughly describe how the terms are used and generally made sense of in everyday language. The overwhelming majority of translators were native speakers of the languages they were translating and, with a few exceptions, two translations per language-country unit were employed to ensure inter-translator reliability.

All translated terms subsequently went through an extensive cleaning process where the translations were reviewed manually using the descriptions provided by the translators. The vast majority of translators were in agreement with each other, and when they provided differing translation suggestions, we consulted machine-learning resources to confirm the most suitable translation. Whereas most terms were suitable for direct translation or could easily be translated into a single term, other terms were more complex. Such terms could refer to several albeit similar concepts, or to a single dense concept that required more contextual understanding for a non-native. Complex terms were therefore substituted with more than one translation and/or a short clarification explaining the essence of the term.

After the translation and cleaning procedure, we continued by carefully clustering terms that pertain to the same concept into topics. More specifically, if a word in lan-

guage X was translated to "nation" and/or "nation-state", and a word in language Y was translated to "state", and a word Y was translated to "nation" and/or "republic" and/or "state", we clustered these words under the umbrella topic "nation, nation-state, republic, state" because they broadly refer to the similar thing. In this process, 3345 translations were reduced to 808 topics telling us how the word democracy is used in online text data across the world at an aggregated level.

3.3 Model and data limitations

As with all word2vec models, CBOW is a Big Data tool and underperforms on small data sizes and infrequent words, particularly if used on a data size smaller than one million words. Table 1 shows descriptive statistics of the data size of all our models by source type measured in number of words. From the standard deviations, it is evident that the text size differs substantially between the models and some models simply lack the sufficient amount of data to produce a satisfactory output.

It is important to note that the sampling process of the CBOW model is randomized and based on cumulative data samples dating back to 2015 when the lexicon was initially built. Since then, new models have been built and all previous models have been updated several times, generating more data and new languages and countries to the lexicon. Because the data for this paper was collected in 2017, it is based on text samples from early 2015 to late 2017. While the randomized text sampling makes the model output fairly resistant to dramatic fluctuations with subsequent updates, the cumulative aspect nonetheless makes over-time changes in the text content impossible to control for. This remains a limitation for the study as such.

A caveat need mentioning is the classification of online text sources where the labels news data and social data should be considered as rather broad source types. First, the data only includes sources that are open to the public. Domains that are closed for commercial and/or privacy reasons are not included in our sample. This means that social platforms such as Facebook or Twitter are inaccessible and for this reason our social data largely consists of web documents from blogs and discussion forums. Second, data vendors do not necessarily have a standard practice for classifying web documents and their practices are not always transparent. While we for instance are well aware that news sites produce content that cannot strictly be defined as news only, it is practically impossible to assess whether news documents containing reader comments are being labeled as news data or not.

We currently have limited capacity to control the content validity of all source texts sampled by the models. The vast amounts of web documents sampled makes it virtually impossible to examine all data manually. Not only would such a daunting

Table 1 Summary Statistics of All DSM Models in Number of Words by Online Source Type (The summary statistics was exported from our updated models in 2019.)

Source	Mean	SD	Min	Max	Total
News	150,200,000	101,644,149	1941	351,100,000	21,175,792,758
Social	143,100,000	102,835,088	17,000	378,700,000	20,181,014,528

task require an intimate knowledge of all languages modeled, familiarity with the internet culture in the given country would be equally important. On top of this, it is not entirely obvious what constitutes representativity online; are news articles with countless page views more representative than rarely visited blog posts? If so, who does news articles represent; the producers of news or the consumers or news? Although there are services that provide certain data on top domains across the world, demographic variables are generally scarce, making the use of such data complicated from a perspective of population-based representation.

Calling for more transparency into the practices of large-scale data vendors may certainly be warranted, but it is worth noting that copyright as well as strengthened legal frameworks for privacy and personal data protection—like the *EU General Data Protection Regulation (GDPR)* (EU 2016/679)—also prevents many vendors from releasing their source data, which ultimately affects the replicability insofar as it makes an exact replication of the models underpinning the lexicon challenging.

The somewhat opaque nature of the data, coupled with the relative variation of available text sizes, obviously raises important methodological issues, particularly where the representativeness of the data is concerned. Using text data that is mainly crawled for commercial purposes means that we by no means can guarantee the same population representation required by comparative survey methodology. Still, vast amounts of language data required to build a lexicon of the same scope and magnitude as ours cannot simply be gathered without the technical, financial and legal resources of a large-scale international enterprise. This trade-off is difficult to circumvent, but because we are interested in how democracy is used in everyday language, this approach seems like the most viable option, even considering the problems with using existing data provider.

4 Findings

4.1 Democracy topics across the world

Fig. 4 shows the 15 most frequent clustered terms or "topics" that are being used in the same way as the word democracy in online news and social media. To begin with, the results evidently have some face validity as they all related to the word democracy in some sense, and several appears to be related to systems of rule. Sovereignty appears to be one of the most frequently mentioned topics related to democracy, along with terms such as nation, republic or state, and community, society or societies. Terms related to autocracy or dictatorship also appear to be very common, particularly in social media. Even though such terms must be considered as quite distant to democracy, their presence in the vector-space is not surprising given that the algorithms provide us with terms used in a similar fashion. This means that antonyms to democracy can also be captured and in this particular case autocracy and dictatorship can be seen as alternative systems of rule to democracy.

The rank correlation between news media and social media for all clustered terms is 0.36 (Spearman's rho). Given the unstructured—and somewhat messy—nature of the data this is rather impressive. Although we largely define news media as different

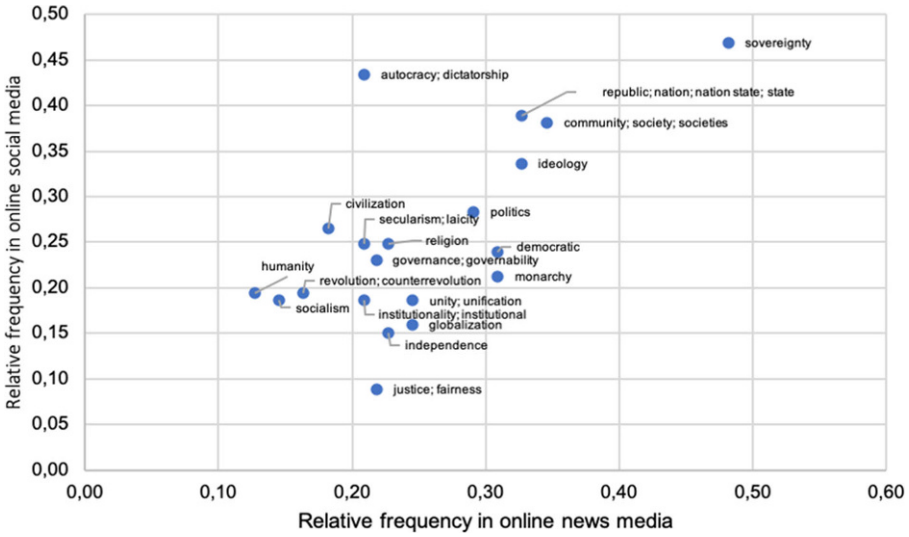


Fig. 4 Top 15 Most Frequent Democracy Topics Across 114 Language-Country Units (Relative frequency is calculated as the absolute frequency divided by the number of existing language-country units from news media and social media.)

types of editorial text published for internet consumers to read, and social media as different types of text produced by internet consumers themselves, we cannot fully guarantee that this is the case. Because of this, we combine the two source types in the subsequent analysis.

In order to make the results more comprehensible and the qualitative interpretation more transparent, we have further grouped the countries included in our data into global regions with similar cultural and historical features based on a division developed by Welzel (2013). This allows for an illustrative aggregation of the most frequent clustered terms across the world. Table 2 shows the top ten most frequently occurring topics across ten regions when we have combined news data and social data (for a more detailed overview of the ten most frequent topics in news media and social media respectively, see Table 4 in the appendix).

Zooming in on regional similarities and differences in Table 2, we find interesting overlaps in language use between different regions. In Catholic and Protestant Europe, the English West and Latin America, we find two topics that occur within all four regions, namely sovereignty, and community and/or society. Across these regions, with our language data covering 55 countries, we find that democracy is generally spoken about in terms of independent nation states where community and sovereignty are critical constituents, as in Robert Dahl's model of polyarchy (1989).

Comparing Europe specifically, the Catholic and Protestant regions share similar notions of democracy, particularly where references to legal justice and the rule of law are concerned, which in Easton's (1975) model refers to specific support. The rule of law, which is an important constituent factor in the quality of government, has also been pointed out as an important factor behind citizens' understanding of or support for democracy in a West European country setting by Dahlberg and Holm-

Table 2 Top Ten Most Frequent Democracy Topics in Combined Online Media Across Ten Global Regions. (Global regions developed by Welzel 2013)

English West	*Protestant Europe*	*Catholic Europe*	*Ex-Communist East*	*Ex-Communist West*
Sovereignty	**Legal state; state of justice; rule of law**	Republic; nation; nation state; state	**Ideology**	**Autocracy; dictatorship**
capitalism	**politics**	**sovereignty**	republic; nation; nation state; state	**ideology**
socialism	**community; society; societies**	**monarchy**	**civilization**	political right-wing
unity; unification	democracy (alt declensions)	**community; society; societies**	elite	political left-wing
nationalism	socialism	**politics**	**religion**	**religion**
secularism; laicity	**autocracy; dictatorship**	secularism; laicity	**autocracy; dictatorship**	authority; power; force; control
community; society; societies	equality; equity	**globalization**	diplomacy	democracy (alt declension)
monarchy	freedom of speech	civilization	bureaucracy	propaganda
multiculturalism	justice; fairness	**autocracy; dictatorship**	political power; state power; superpower	christianity
liberalism	**sovereignty**	community; fellowship; brotherhood	revolution; counterrevolution	**civilization**
		freedom		legal state; state of justice; rule of law
		legal state; state of justice; rule of law		oligarchy
		religion		political
		social democracy		
		tolerance		

berg (2014). One striking difference is the emphasis on individual political values in the Protestant Europe, where the data is more closely connected to different forms of freedoms and rights such as equality, freedom of speech, justice and fairness. In contrast, the Catholic counterpart is more focused on the community with topics related to civilization, republic and nation, fellowship and brotherhood as well as tolerance and social democracy.

Most topics in the English West pertain to different forms of political ideologies such as socialism, liberalism and nationalism but also to other forms of belief systems such as capitalism, multiculturalism and secularism. Apparently, democracy in the English Western hemisphere is spoken about in terms of an ideational system pertaining to values and ideas relating to the input side of the democratic system, if we are relating it back to the Eastonian model of democracy. Unsurprisingly, some topics in Latin America resembles those of Catholic Europe, but the Latin American data contains more references to different system-related features such as institutionality, legality and governance. Latin America and the former communist regions also share some similar notions of democracy, namely oligarchy and revolution.

Notable similarities can be identified between Ex-Communist East and Ex-Communist West, particularly with topics such as ideology, religion and civilization. Power also appears to be a common denominator; in the East it is referred to as more specific political state power or superpower, while the West has more general references to power in terms of authority, control or force. While language data from Ex-Communist West has a stronger emphasis on specific political wings, democracy in Ex-Communist East is talked about in terms of a political or social elite, which may be reflective of the fact that a larger share of the countries in Ex-Commu-

Table 2 (Continued)

Latin America	*North Africa and the Middle East*	*Africa*	*East Asia*	*South Asia*
Institutionality; institutional **sovereignty** **community; society; societies** governance; governability **globalization** oligarchy **monarchy** revolution; counterrevolution tyranny legality	Secularism; laicity religion citizenship **peace** **cohesion** dignity independence **democratic** modernity republic; nation; nation state; state	**Democratic** ethics; ideals; principles; values unity; unification governance; governability **peace** **sovereignty** **autocracy; dictatorship** politics stability **cohesion**	**Capitalism** **democratic** **unity; unification** republic; nation; nation state; state independence **community; society; societies** **autocracy; dictatorship** ideology self-determination socialist	**Democratic** **unity; unification** diversity ideology **sovereignty** **capitalism** politics secular pluralism; pluralistic socialism

Relative frequency is calculated as the absolute frequency divided by the number of existing language-country units from news media and social media per region

Catholic Europe Andorra, Austria, Belgium, Cyprus, France, Greece, Israel, Italy, Luxembourg, Portugal, Spain

Protestant Europe Denmark, Finland, Germany, the Netherlands, Norway, Sweden, Switzerland

English West Australia, Canada, Ireland, New Zealand, United Kingdom, United States

Latin America Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Uruguay, Venezuela

Ex-Communist East Armenia, Azerbaijan, Belarus, Bulgaria, Kazakhstan, Kyrgyzstan, Moldova, Romania, Russia, Ukraine

Ex-Communist West Czech Republic, Hungary, Latvia, Lithuania, Poland, Slovakia, Slovenia

East Asia China, Hong Kong, Taiwan, Thailand, Vietnam

South Asia Brunei Darussalam, India, Indonesia, Malaysia, Pakistan, Philippines, Singapore, Sri Lanka

North Africa and the Middle East Algeria, Egypt, Iran, Morocco, Tunisia, Turkey

Sub-Saharan Africa Cameroon, Gabon, Ghana, Kenya, Mali, Namibia, Nigeria, Senegal, Somalia, South Africa, Zambia, Zimbabwe

NB #1 In the original classification, Armenia, Azerbaijan, Kazakhstan and Kyrgyzstan belong to Central Asia. Because our data do not cover any other countries from Central Asia and these countries were affiliated with the former Soviet Union, we have included them in the region Ex-Communist East

NB #2 For Catholic Europe and Ex-Communist West, more than ten topics yield the same frequency score and are consequently included in the top ten most frequent topics

NB #3 Most frequent topics across all regions are marked in bold

nist West are representative democracies with more room for political contest and a wider variety of political actors generally.

Turning to regions in Africa, the Middle East and Asia, we find even more interesting overlaps in language use. In North Africa and the Middle East as well as in Sub-Saharan Africa, we find that topics related to cohesion and peace tend to co-occur more frequently than others. Similarly, unity and unification appear in Sub-Saharan Africa just as in East and South Asia. Such topics—coupled with topics such as citizenship, independence, self-determination, sovereignty and stability across these same regions—are prerequisites for state-building, which is telling as many

countries in this region are marked by a violent colonial past or more recent civil conflict where democracy becomes synonymous with desirable political outcomes. Importantly, the results from these regions also point to the fact that democratic principles and values, so called input factors of the political system, are indeed a topic of conversation online in regions with more newly established or fragile democracies, which is supported by the findings from Bratton and Mattes (2001).

4.2 Classification of democracy topics

The theoretical attempts to portray people's conceptions of democracy, as laid out by Easton (1975) and Norris (1999), have gained validity in a number of empirical correlational studies (see Bratton and Mattes 2001; Aarts and Thomassen 2008; Dahlberg, Linde & Holmberg 2015). Inspired by their conceptualizations of support for democracy, we have constructed a scheme that roughly based on the separation between diffuse and specific support. Although Easton (1975) and Norris (1999) specifically are studying support for democracy it is still, we argue, a valid conceptualization of different aspects of the democratic system at different levels, which fruitfully can be used for classifying and thus studying meaning differences. From this distinction follows that the separation not only is a matter of different levels of abstraction but also a difference in terms of input and output of the democratic system (it should be underlined that our classification scheme is partly deductive based on by Easton (1975) and Norris (1999) and partly inductive, based on the empirical data). If we are able to conceptualize language use for the term democracy into a smaller set of theoretically meaningful categories for different languages; we may also be able to incorporate the proportions of stances for each language within each category back to the survey-based data at a later stage. For instance, such language-based variable constructs could potentially be used to control for differences in meaning for the word democracy across languages. For this paper, however, we only apply the scheme to the most frequent democracy topics found in our data.

Table 3 displays the classification matrix with five categories ranging from more diffuse, at the top, to more specific levels of abstraction at the bottom. The first category (*Community*) refers to democracy topics pertaining to the political community more broadly, or the collective society in which the political system is situated—e.g. community or civilization. The second category (*Belief*) contains topics that refer to systems of belief including ideology and religion, doctrine, or other beliefs or ideas that make up the political system—e.g. Christianity or liberalism. The third category (*Principle*) is an input category where we find democracy topics that relate

Table 3 Classification Matrix

Level of Abstraction	Thematic Category	Classification Code
Diffuse	Community	1
	Belief	2
Specific	Principle	3
	Governance	4
	Outcome	5

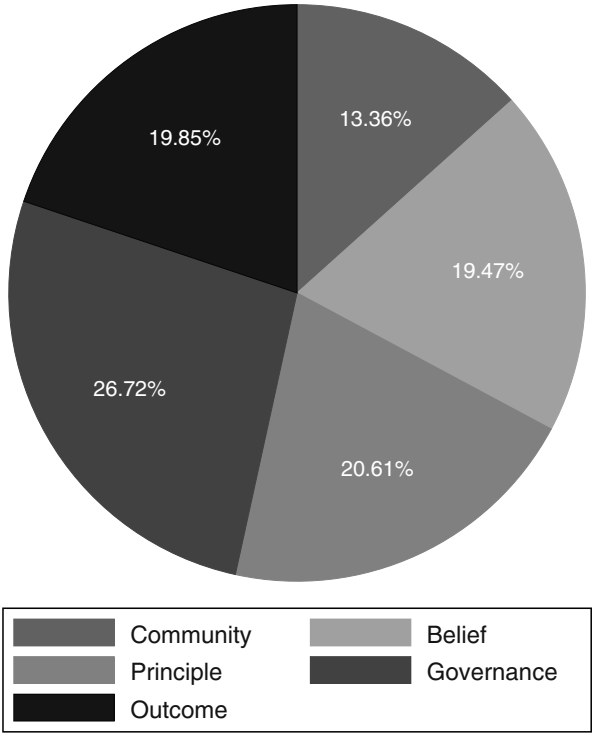


Fig. 5 Classification of Most Frequent Democracy Topics Across 114 Language-Country Units

to principles, norms and values—e.g. freedom of speech or sovereignty. Category 4 (*Governance*) is considered a throughput category and consists of topics related to governance, from systems of rule to system-related procedures—e.g. autocracy or politics. The last category (*Outcome*) includes topics that are ultimately refers to outputs or outcomes of the political system—e.g. welfare state or peace.

The first results of the classification are available in Fig. 5 where news media and social media have been combined in order to get an overall picture of the meaning of democracy. The most frequent category is governance followed by principles, belief and outcome. That governance is in top is not surprising since democracy is governance, and the least common denominator of the democracy concept in political theory is about the rule by the people (Dahl 1989). More interestingly though is the fact that it is aspects of democracy at the lower levels of abstraction that are most commonly occurring in language media data, a result that correspond well with what we know about the drivers of support for democracy. From prior research we know that one of the most frequently used survey measures on support for democracy (satisfaction with the way democracy works), mainly is a performance measure correlated with factors found at the more specific level of democracy (Linde and Ekman 2003; Dahlberg et al. 2015). This is an encouraging finding since meaning differences or rather dimensions of democracy, identified in online text data, corresponds to what we know about the concept from survey research data.

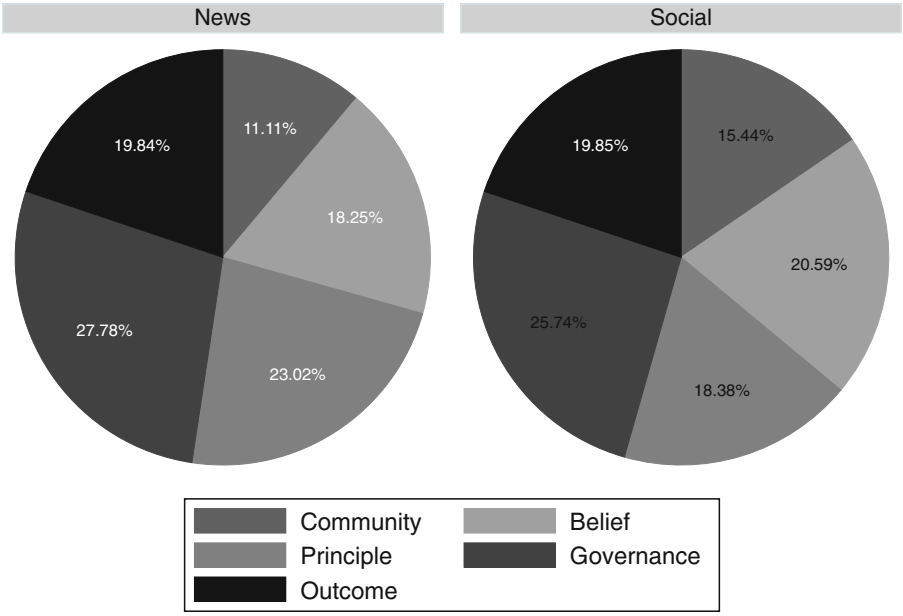


Fig. 6 Classification of Most Frequent Democracy Topics by Source Type

In the case of robustness, we continue to separate the partly theoretical classification between different media types. Fig. 6 shows the classification results from news media and social media respectively.

The results in Fig. 6 do not reveal any major differences between the two media types. To some extent, the community category is somewhat larger in social media compared to news media, mainly on behalf of the category labeled principles. This difference does have face validity in the sense that principles in terms of freedom of speech is more outspoken in news media data, while terms related to the society and the people is more commonly occurring in social media. If one would have to guess, we believe that this is what one should expect. Journalists are more focused on democracy as values while the "people" is more concerned about democracy in terms of identity and society (if we believe that social media is a better proxy for meaning among some kind of a population separated from news media and journalists). To what extent some media types are more representative to some kind of a population in a similar manner as survey data is another discussion far beyond the scope of this study.

The major overlaps between news media and social media, however, makes us more confident that we for now can collapse the two media categories. Fig. 7 shows the classification results across the ten global regions. More detailed and media-separated numbers are available in Table 5 in the appendix.

The partition of our classification across different cultural regions, based on a division made by Welzel (2013) using World Values Survey data, reveals some interesting patterns. For instance, we find that governance, which was the largest category overall, varies substantially between regions and makes up about 50% of the terms

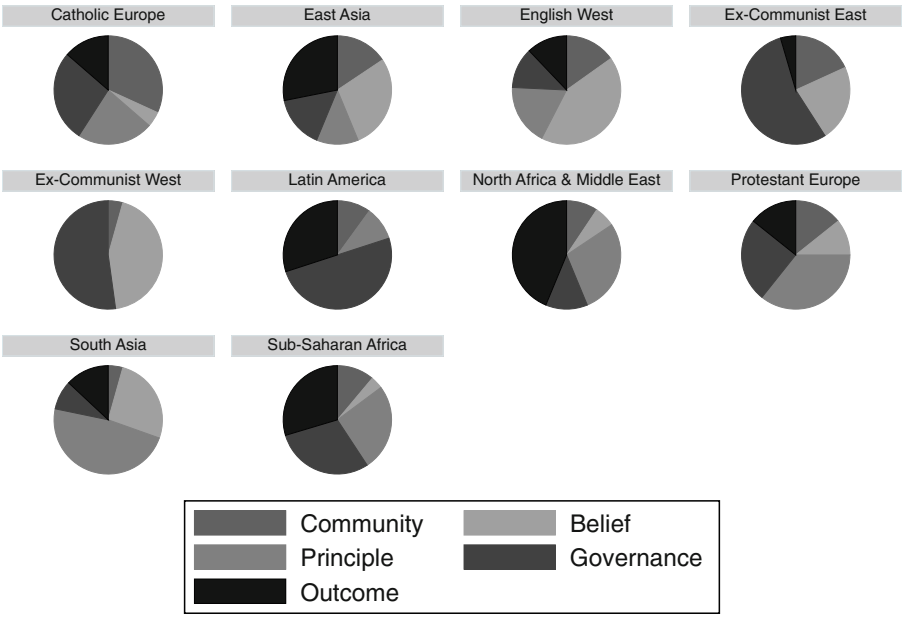


Fig. 7 Classification of Most Frequent Democracy Topics Across Ten Global Regions

in former communist countries and Latin America while being less pronounced in particularly South Asia, where it instead is democracy as principles that dominates. Democracy as outcomes is less pronounced in Europa as well as the English West while making up one of the larger categories in North Africa and the Middle East, Sub-Saharan Africa, Latin America and East Asia. This could be an indication on that conceptions of democracy also are affected by democratic consolidation, that is how long a country has been a democracy, as suggested by Dahlberg et al. (2015). The more diffuse conceptions of democracy such as community and belief system are more equally distributed. Democracy in terms of community is largest in Catholic Europe while belief system is most pronounced in Ex-Communist and English West. In the latter examples, it is mainly democracy spoken about in the same terms as other ideological belief systems such as capitalism, liberalism or socialism. The fact that the term democracy is occurring in similar language contexts as the large political ideologies is intriguing. It can be interpreted as that democracy is conceptualized as something more than a steering system. In these parts of the world it is also an overarching ideology with values and ideas.

The classification scheme used in this paper has mainly been developed for an illustrative purpose and can and should be developed much further. However, it shows how natural language processing techniques applied on large online text data can yield patterns and results that corresponds to what we knew from comparative survey-based research.

5 Final remarks

In this paper, we have taken some first steps in attempting to capture variations in the meaning of and usage of democracy across the globe. We have done so by combining news and social online text data onto which we have applied a natural language processing technique known as distributional semantics, used to capture structure and meaning in language use. The qualitative interpretations of the distributional semantic models indicate that our data and our methodological approach for analyzing the data has validity. Although the picture is painted with wide brush strokes systematic pattern appears.

Exploring the most frequently occurring clustered terms or topics that are being used in the same way as the word democracy, we find that data from Europe and the English Western world reflect meanings of democracy that is primarily related to the input side of the democratic system (Easton 1975). We can identify important cultural differences between the regions where the most frequent topics from the Catholic Europe have a larger focus on community vis-à-vis the Protestant Europe, where individual political freedoms and rights are more pronounced. In turn, the topics from the English West is more reflective of different political ideologies than collective or individual principles and values. Though Latin America and the Catholic Europe share similar notions of democracy, the former region contains more procedural topics pertaining to conditions of rule such as governance, legality and institutionality. Language data from Ex-Communist East and West further tells us that democracy is used synonymously with ideology, religion, political affiliation and power in these regions. In Africa and Asia, we find topics pertaining to state-building where independence, stability, unity and peace are used in the same fashion as democracy. Such representations are important conditions for democracy, but they can also be considered as vital outcomes of a democratic system.

Ideally, our ambition has been to contribute to the field of country comparative survey research by providing a flexible tool for capturing meaning differences across linguistically and culturally diverse nations. The results from this paper shows that there are some common understandings of the concept of democracy but that there also are regional variations that the comparative surveys are unable to capture. Having said that, we do not pretend that we are able to grasp the various understandings of democracy in a population representative manner. Rather, we are pinning down conceptualizations from the public debate where it occurs in the 21st century, in not only news data but also in social data. From a survey perspective, this is probably a good enough proxy for the common understanding of more abstract concepts such as democracy. Although we might not have solved the issue comparability, we have at least provided some further knowledge on how a concept like democracy travels across languages and cultures—knowledge that can be used for designing surveys in the future and to analyze public support for democracy.

Some venues for further research should be to place a stronger emphasis on the validity, reliability and representativity of online data. In this paper, we find clear signals between the findings of NLP analyses of online text data versus previous analyses of population-based survey data. This is encouraging and in line with our expectations as text data, just as a survey response, is not produced in a vacuum.

An important question to be further explored is what are the relevant dimensions of representativity of online text data in relation to survey data? What individual, linguistic, cultural and institutional characteristics define different online landscapes across the world, and what are the implications of using online text data as a complement to survey data? Representativity online remains an open question as much as a gap in the existing literature. Whilst we are currently involved in related research aimed at estimating demographic representation in online text data, we can only conclude that such estimations are beyond the scope of this paper, and we hope this paper can encourage other researchers in the field will continue to explore the under-researched intersection of computational linguistics and political science.

Acknowledgements The authors would like to extend their utmost gratitude the research assistants that provided essential translation services for the purpose of this study. We are also grateful to our colleagues at the interdisciplinary research program Linguistic Explorations of Societies, particularly Amaru Cuba Gyllensten, Ariel Ekgren and Magnus Sahlgren at the Research Institutes of Sweden (RISE). This study would be impossible without the funding received from the Swedish Research Council.

Funding This study was funded by the Swedish Research Council [grant number 421-2014-1393].

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Sofia Axelsson and Stefan Dahlberg. The first draft of the manuscript was written by Stefan Dahlberg and Sofia Axelsson and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by University of Gothenburg.

Conflict of interest S. Dahlberg, S. Axelsson and S. Holmberg declare that they have no competing interests.

Appendix

Table 4 Top Ten Most Frequent Democracy Topics Across Ten Global Regions. (Global regions developed by Welzel 2013)

Cultural Region	News media	Frequency	Code	Social media	Frequency	Code
Catholic Europe	Republic; nation; nation state; state	0.60	1	Republic; nation; nation state; state	0.67	1
	Sovereignty	0.60	3	Sovereignty	0.67	3
	Community; society; societies	0.40	1	Autocracy; dictatorship	0.47	4
	Globalization	0.40	5	Monarchy	0.47	4
	Monarchy	0.40	4	Community; society; societies	0.40	1
	Politics	0.40	4	Politics	0.40	4
	Secularism; laicity	0.40	3	Secularism; laicity	0.40	3
	Civilization	0.33	1	Civilization	0.33	1
	Justice; fairness	0.33	5	Community; fellowship; brotherhood	0.33	1
	Legal state; state of justice; rule of law	0.33	4	Globalization	0.33	5
	Tolerance	0.33	3	Religion	0.33	2
Protestant Europe	Community; fellowship; brotherhood	0.64	1	Autocracy; dictatorship	0.64	4
	Legal state; state of justice; rule of law	0.64	4	Legal state; state of justice; rule of law	0.64	4
	Justice; fairness	0.55	5	Community; society; societies	0.55	1
	Politics	0.55	4	Politics	0.55	4
	Community; society; societies	0.45	1	Socialism	0.55	2
	Democracy (alt. declension)	0.45	4	Democracy (alt. declension)	0.45	4
	Equality; equity	0.45	3	Ideology	0.36	2
	Freedom	0.45	3	Market economy	0.36	5
	Freedom of speech	0.45	3	Welfare society; welfare state	0.36	5
	Sovereignty	0.45	3	Capitalism	0.27	2
	–	–	–	Democratic	0.27	3
	–	–	–	Equality; equity	0.27	3
	–	–	–	Freedom of speech	0.27	3
	–	–	–	Globalization	0.27	5
	–	–	–	Neutrality	0.27	3
	–	–	–	Republic; nation; nation state; state	0.27	1
	–	–	–	Secularism; laicity	0.27	3
	–	–	–	Sovereignty	0.27	3

Table 4 (Continued)

Cultural Region	News media	Frequency	Code	Social media	Frequency	Code
English West	Sovereignty	1.00	3	Secularism; laicity	0.88	3
	Capitalism	0.75	2	Sovereignty	0.88	3
	Monarchy	0.75	4	Capitalism	0.75	2
	Socialism	0.75	2	Community; society; societies	0.75	1
	Unity; unification	0.75	5	Nationalism	0.75	2
	Nationalism	0.63	2	Socialism	0.75	2
	Politics	0.63	4	Unity; unification	0.75	5
	Multiculturalism	0.50	2	Liberalism	0.63	2
	Community; society; societies	0.38	1	Multiculturalism	0.63	2
	Conservatism	0.38	2	Autocracy; dictatorship	0.38	4
	Democratic	0.38	3	Civilization	0.38	1
	Ideology	0.38	2	Conservatism	0.38	2
	Imperialism	0.38	5	Fascism	0.38	2
	Independence	0.38	5	Humanity	0.38	1
	Liberalism	0.38	2	Individualism	0.38	2
	Secularism; laicity	0.38	3	Monarchy	0.38	4
	–	–	–	Republic; nation; nation state; state	0.38	1
Latin America	Institutionality; institutional	1.00	4	Institutionality; institutional	0.90	4
	Sovereignty	0.90	3	Sovereignty	0.90	3
	Community; society; societies	0.81	1	Community; society; societies	0.71	1
	Monarchy	0.71	4	Governance; governability	0.67	4
	Globalization	0.62	5	Fight; (social) struggle	0.48	5
	Governance; governability	0.62	4	Revolution; counterrevolution	0.48	5
	Oligarchy	0.62	4	Globalization	0.43	5
	Tyranny	0.62	5	Oligarchy	0.43	4
	Constitution	0.43	4	Citizenship	0.33	5
	Legality	0.43	4	Monarchy	0.29	4

Table 4 (Continued)

Cultural Region	News media	Frequency	Code	Social media	Frequency	Code
Ex-Communist East	Ideology	0.92	2	Ideology	0.77	2
	Elite	0.85	4	Republic; nation; nation state; state	0.77	1
	Republic; nation; nation state; state	0.85	1	Civilization	0.69	1
	Civilization	0.69	1	Autocracy; dictatorship	0.62	4
	Diplomacy	0.69	4	Religion	0.62	2
	Religion	0.62	2	Bureaucracy	0.38	4
	Autocracy; dictatorship	0.54	4	Diplomacy	0.38	4
	Political power; state power; superpower	0.54	4	Elite	0.38	4
	Revolution; counter-revolution	0.54	5	Democracy (alt. declension)	0.31	4
	Bureaucracy	0.46	4	Doctrine	0.31	2
	Censorship	0.46	4	Journalism	0.31	4
Ex-Communist West	Autocracy; dictatorship	0.86	4	Autocracy; dictatorship	0.86	4
	Ideology	0.71	2	Ideology	0.71	2
	Political right-wing	0.71	2	Civilization	0.57	1
	Authority; power; force; control	0.57	4	Political left-wing	0.57	2
	Political	0.57	4	Political right-wing	0.57	2
	Religion	0.57	2	Democracy (alt. declension)	0.43	4
	Christianity	0.43	2	Oligarchy	0.43	4
	Democracy (alt. declension)	0.43	4	Propaganda	0.43	4
	Diplomacy	0.43	4	Religion	0.43	2
	Legal state; state of justice; rule of law	0.43	4	Rhetoric	0.43	3
	Liberalism	0.43	2	–	–	–
	Political left-wing	0.43	2	–	–	–
	Propaganda	0.43	4	–	–	–

Table 4 (Continued)

Cultural Region	News media	Frequency	Code	Social media	Frequency	Code
East Asia	Unity; unification	0.83	5	Autocracy; dictatorship	0.50	4
	Capitalism	0.50	3	Politics	0.50	4
	Democratic	0.50	3	Republic; nation; nation state; state	0.50	1
	Ideology	0.50	2	Capitalism	0.38	2
	Independence	0.50	5	Communism	0.38	2
	Leadership; leader; leaders	0.50	4	Community; society; societies	0.38	1
	Self-determination	0.50	5	Democratic	0.38	3
	Community; society; societies	0.33	1	Equality; equity	0.38	3
	Islam; islamic	0.33	2	Legal state; state of justice; rule of law	0.38	4
	Patriotism; patriotic	0.33	2	(Non-violent) resistance; civil disobedience; protest	0.25	5
	Peaceful; peacefulness	0.33	3	Economic system; economy; economics	0.25	5
	Republic; nation; nation state; state	0.33	1	Human rights	0.25	5
	Rob; seize; steal	0.33	5	Humanity	0.25	1
	Socialist	0.33	2	Independence	0.25	5
	–	–	–	Institution; system	0.25	4
	–	–	–	Prosperity	0.25	5
	–	–	–	Socialism	0.25	2
	–	–	–	Socialist	0.25	2
South Asia	Unity; unification	0.90	5	Democratic	0.70	3
	Democratic	0.80	3	Ideology	0.70	2
	Sovereignty	0.70	3	Capitalism	0.60	2
	Diversity	0.60	3	Diversity	0.50	3
	Politics	0.60	4	Unity; unification	0.50	5
	Peace	0.50	5	Humanity	0.40	1
	Pluralism; pluralistic	0.50	3	Politics	0.40	4
	Secular	0.50	3	Secular	0.40	3
	Capitalism	0.40	2	Secularism; laicity	0.40	3
	Ethics; ideals; principles; values	0.40	3	Socialism	0.40	2
	Ideology	0.40	2	Sovereignty	0.40	3
	Socialism	0.40	2	–	–	–

Table 4 (Continued)

Cultural Region	News media	Frequency	Code	Social media	Frequency	Code
North Africa and the Middle East	Citizenship	0.67	5	Religion	0.71	2
	Peace	0.67	5	Secularism; laicity	0.71	3
	Secularism; laicity	0.67	3	Citizenship	0.43	5
	Cohesion	0.50	5	Cohesion	0.43	5
	Democratic	0.50	3	Dignity	0.43	3
	Dignity	0.50	3	Independence	0.43	5
	Independence	0.50	5	Legitimacy	0.43	5
	Religion	0.50	2	Modernity	0.43	5
	Decentralization	0.33	4	Peace	0.43	5
	Democracy (alt. declension)	0.33	4	Politics	0.43	4
	Freedom	0.33	3	Republic; nation; nation state; state	0.43	1
	Justice; fairness	0.33	5	Sovereignty	0.43	3
	Moderation	0.33	4	Stability	0.43	5
	Modernity	0.33	5	Tolerance	0.43	3
	Republic; nation; nation state; state	0.33	1	–	–	–
	Thought; reflection	0.33	5	–	–	–
	Tolerance	0.33	3	–	–	–
	Will; willingness; volition	0.33	1	–	–	–

Table 4 (Continued)

Cultural Region	News media	Frequency	Code	Social media	Frequency	Code
Sub-Saharan Africa	Democratic	0.62	3	Democratic	0.62	3
	Ethics; ideals; principles; values	0.62	3	Sovereignty	0.54	3
	Governance; governability	0.54	4	Autocracy; dictatorship	0.46	4
	Peace	0.46	5	Cohesion	0.46	5
	Politics	0.46	4	Stability	0.46	5
	Unity; unification	0.46	5	Unity; unification	0.46	5
	Autocracy; dictatorship	0.31	3	Community; society; societies	0.38	1
	Equality; equity	0.31	3	Constitution	0.38	4
	Republic; nation; nation state; state	0.31	1	Peace	0.38	5
	Sovereignty	0.31	3	Ethics; ideals; principles; values	0.31	3
	Stability	0.31	5	Governance; governability	0.31	4
	–	–	–	Leadership; leader; leaders	0.31	2
	–	–	–	Nationalism	0.31	4
	–	–	–	Politics	0.31	2
	–	–	–	Republic; nation; nation state; state	0.31	4
	–	–	–	Revolution; counterrevolution	0.31	5

For country classification, see Table 2
Frequency represents the relative frequency calculated as the absolute frequency divided by the number of existing language-country units from news media and social media per region
Code represents the classification of each topic described in Table 3

Table 5 Classification of Most Frequent Democracy Topics Across Global Regions

Classification by Global Regions	News Media (%)	Social Media (%)	Total (%)	Difference (%)
Catholic Europe				
Community	27.3	36.4	31.8	–9.1
Belief	0.0	9.1	4.6	–9.1
Principle	27.3	18.2	22.7	9.1
Governance	27.3	27.3	27.3	0.0
Outcome	18.2	9.1	13.6	9.1
Protestant Europe				
Community	20.0	11.1	14.3	8.9
Belief	0.0	16.7	10.7	–16.7
Principle	40.0	33.3	35.7	6.7
Governance	30.0	22.2	25.0	7.8
Outcome	10.0	16.7	14.3	–6.7
English West				
Community	6.3	23.5	15.2	–17.3
Belief	37.5	47.1	42.4	–9.6
Principle	25.0	11.8	18.2	13.2
Governance	12.5	11.8	12.1	0.7
Outcome	18.8	5.9	12.1	12.9
Latin America				
Community	10	10	10	0.0
Belief	0.0	0.0	0.0	0.0
Principle	10	10	10	0.0
Governance	60	40	50	20
Outcome	20	40	30	–20
Ex-Communist East				
Community	18.2	18.2	18.2	0.0
Belief	18.2	27.3	22.7	–9.1
Principle	0.0	0.0	0.0	0.0
Governance	54.6	54.6	54.6	0.0
Outcome	9.1	0.0	4.6	9.1
Ex-Communist West				
Community	0.0	10.0	4.4	–10.0
Belief	46.2	40.0	43.5	6.2
Principle	0.0	0.0	0.0	0.0
Governance	53.9	50.0	52.2	3.9
Outcome	0.0	0.0	0.0	0.0
East Asia				
Community	14.3	16.7	15.6	–2.4
Belief	35.7	22.2	28.1	13.5
Principle	14.3	11.1	12.5	3.2
Governance	7.1	22.2	15.6	–15.1
Outcome	28.6	27.8	28.1	0.8

Table 5 (Continued)

Classification by Global Regions	News Media (%)	Social Media (%)	Total (%)	Difference (%)
South Asia				
Community	0.0	9.1	4.4	–9.1
Belief	25.0	27.3	26.1	–2.3
Principle	50.0	45.5	47.8	4.6
Governance	8.3	9.1	8.7	–0.8
Outcome	16.7	9.1	13.0	7.6
North Africa and the Middle East				
Community	11.1	7.1	9.4	4.0
Belief	5.6	7.1	6.3	–1.6
Principle	27.8	28.6	28.1	–0.8
Governance	16.7	7.1	12.5	9.5
Outcome	38.9	50.0	43.8	–11.1
Sub-Saharan Africa				
Community	9.1	12.5	11.1	–3.4
Belief	0.0	6.3	3.7	–6.3
Principle	36.4	18.8	25.9	17.6
Governance	27.3	31.3	29.6	–4.0
Outcome	27.3	31.3	29.6	–4.0

References

Aarts, Kees, and Jacques Thomassen. 2008. Satisfaction with democracy: do institutions matter? *Electoral Studies* 27:5–18.

Antoniak, Maria, and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics* 6:107–119.

Ariely, Gal. 2014. Deocracy-assessment in cross-national surveys: a critical examination of how people evaluate their regime. *Social Indicators Research* 121:621–635.

Behling, Orlando, and Kenneth S. Law. 2000. *Translating questionnaires and other research instruments: Problems and solutions*. Thousand Oaks: SAGE.

Bratton, Michael, and Robert Mattes. 2001. Support for democracy in africa: intrinsic or instrumental? *British Journal of Political Science* 31(3):447–474.

Braun, Michael, and Janet Harkness. 2005. Text and context: challenges to comparability in survey questions. In *Methodological aspects in cross-national research*, ed. Jürgen H.P. Hoffmeyer-Zlotnok, Janet Harkness, 95–107. Mannheim: GESIS-ZUMA.

Canache, Damarys, Jeffery J. Mondak, and Mitchell A. Seligson. 2001. Meaning and measurement in cross-national research on satisfaction with democracy. *Public Opinion Quarterly* 65:506–528.

Cutler, Fred, Andrea Nuesser, and Ben Nyblade. 2013. *Evaluating the quality of democracy with individual level models of satisfaction: or, a complete model of satisfaction with democracy*. ECPR General Conference, Bordeaux, 4–7 September 2013.

Dahl, Robert A. 1989. *Democracy and its critics*. New Haven: Yale University Press.

Dahlberg, Stefan, and Sören Holmberg. 2014. Democracy and bureaucracy: how their quality matters for popular satisfaction. *West European Politics* 37:515–537.

Dahlberg, Stefan, Jonas Linde, and Sören Holmberg. 2015. Democratic discontent in old and new democracies—Assessing the importance of democratic input and governmental output. *Political Studies* 63:18–37.

Dahlberg, Stefan, Sofia Axelsson, Magnus Sahlgren, Amaru Cuba Gyllensten, and Ariel Ekgren. 2020. *The LES Distributional Semantic Lexicon*. Working Paper. Gothenburg: University of Gothenburg.

Dalton, Russel J. 2004. *Democratic challenges, democratic choices*. Oxford: Oxford University Press.

Dalton, Russell J., Shin Doh-Chull, and Willy Jou. 2007. Understanding democracy: data from unlikely places. *Journal of Democracy* 18(4):142–256.
Diamond, Larry. 1999. *Developing democracy: toward consolidation*. Baltimore: Johns Hopkins University Press.
Easton, David. 1975. A re-assessment of the concept of political support. *British Journal of Political Science* 5(4):435–457.
Firth, John Rupert. 1957. *A synopsis of linguistic theory 1930–1955*. Oxford:: Oxford University Press. Special Volume of the Philological Society.
Foa, Stefan Roberto, and Yascha Mounk. 2016. The danger of deconsolidation: the democratic disconnect. *Journal of Democracy* 27(3):5–17.
Foa, Stefan Roberto, and Yascha Mounk. 2017. The signs of deconsolidation. *Journal of Democracy* 28(1):5–15.
Hoffmeyer-Zlotnok, Jürgen H.P., and Janet Harkness (eds.). 2005. *Methodological aspects in cross-national research*. Mannheim: GESIS-ZUMA.
Holmberg, Sören. 2014. Feeling policy represented. In *Elections and democracy: representation and accountability*, ed. Jacques Thomassen, 132–152. Oxford: Oxford University Press.
Inglehart, Ronald. 2003. How solid is mass support for democracy: And how can we measure it? *Political Science and Politics* 36(1):51–57.
Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, cultural change, and democracy*. Cambridge: Cambridge University Press.
Jeffrey, Pennington, Richard Socher, and Christopher Manning. 2014. *GloVe: global vectors for word representation*. Proceeding of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 10.2014.
Karlgren, Jussi, and Magnus Sahlgren. 2001. From words to understanding. In *Foundations of real-world intelligence*, 294–308. Standford, California: CSLI Publications.
Kasfir, Nelson. 2000. 'Democracy in translation: understanding politics in an unfamiliar culture.': by Frederic C Schaffer. Ithaca, NY: Cornell University Press. *American Political Science Review* 94(3):757–758.
King, Gary Christopher J.L.Murray, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review* 98(1):191–207.
Klingemann, Hans-Dieter. 1999. Mapping political support in the 1990s. In *Critical citizens: global support for democratic governance*, ed. Pippa Norris, 31–56. Oxford: Oxford University Press.
Kriesi, Hanspeter, Leonardo Morlino, Pedro Magalhães, Sonia Alonso, and Mónica Ferrín. 2013. *European Social Survey. Round 6 module on europeans' understandings and evaluations of democracy—Final module in template*. London Centre for Comparative Social Surveys, City University London.
Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica* 20(1):1–31.
Levy, Omer, and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 2, 302–308.
Lindberg, Staffan I., Michael Coppedge, Jogn Gerring, and Jan Teorell. 2014. V-Dem: a new way to measure democracy. *Journal of Democracy* 25(2):159–169.
Linde, Jonas, and Joakim Ekman. 2003. Satisfaction with democracy: a note on a frequently used indicator in comparative politics. *European Journal of Political Science Research* 42(3):391–408.
Lipset, Seymour Martin. 1959. Some social requisites of democracy: economic development and political legitimacy. *American Political Science Review* 53(1):69–105.
Lührmann, Anna, Lisa Gastaldi, Sandra Grahn, I. Lindberg Staffan, Laura Maxwell, Valeriya Mechkova, Richard Morgan, Natalia Stepanova, and Shreeya Pillai. 2019. *V-Dem annual democracy report 2019. Democracy facing global challenges*. Gothenburg: V-Dem Institute, University of Gothenburg.
Lyons, John. 1991. *Natural language and universal grammar*. Cambridge: Cambridge University Press.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.
Mishler, William, and Richard Rose. 2001. Political support for incomplete democracies: realist vs. Idealist theories and measures. *International Political Science Review* 22(4):303–320.
Mounk, Yascha. 2018. *The people vs. Democracy. Why our freedom is in danger and how to save it*. Cambridge: Harvard University Press.
Norris, Pippa (ed.). 1999. *Critical citizens: global support for democratic governance*. Oxford: Oxford University Press.

Norris, Pippa. 2017. Is Western democracy backsliding? Diagnosing the risks. *Journal of Democracy*. https://doi.org/10.2139/SSRN.2933655.

O'Donnell, Guillermo. 2007. The perpetual crises of democracy. *Journal of Democracy* 18(1):5–11.

Pharr, Susan J., and Robert D. Putnam (eds.). 2000. *Disaffected democracies: what's troubling the trilateral democracies?* Princeton: Princeton University Press.

Pierrejean, Bénédicte, and Ludovic Tanguy. 2018. *Predicting word embeddings variability*. Proceedings of the 7th Joint Conference on Lexical and Computational Semantics., 154–159.

Rogers, Anna, Shashwath Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. *What's in your embedding, and how it predicts task performance*. Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, 08.2018.

Sahlgren, Magnus. 2005. *An introduction to random indexing*. Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, Copenhagen, 08.2005.

Sahlgren, Magnus. 2006. *The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm: Department of Linguistics, Stockholm University.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Rivista di Linguistica* 20(1):33–53.

Sahlgren, Magnus, and Alessandro Lenci. 2016. *The effects of data size and frequency range on distributional semantic models*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, 11.2016.

Schaffer, Frederic Charles. 2000. *Democracy in translation: understanding politics in an unfamiliar culture*. New York: Cornell University Press.

Schedler, Andreas, and Rodolfo Sarsfield. 2007. Democrats with adjectives: linking direct and indirect measures of democratic support. *European Journal of Political Research* 46(5):637–659.

Shin, Doh Chull. 2017. Popular understanding of democracy. In *Oxford research Encyclopaedia of politics* https://doi.org/10.1093/acrefore/9780190228637.013.80.

Turney, Peter D., and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188.

Welzel, Christian. 2013. *Freedom rising: human empowerment and the quest for emancipation*. New York: Cambridge University Press.

Welzel, Christian, and Helen Kirsch. 2017. Democracy misunderstood: authoritarian notions of democracy around the globe. *World Values Research* 9(1):1–29.

Welzel, Christian, and Hans-Dieter Klingemann. 2008. Evidencing and explaining democratic congruence: the perspective of "substantive" democracy. *World Values Research, WVR* 1(3):57–90.

Wittgenstein, Ludwig. 1958. *Philosophical investigations*. Oxford: Blackwell.

Zavala-Rojas, Diana, and Willem E. Saris. 2018. Measurementinvariance in multilingual survey research: the role of the language of the questionnaire. *Social Indicators Research* 140(2):485–510.

Řehůřek, Radim, and Petr Sojka. 2010. *Software framework for topic modelling with large corpora*. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, 05.2010.