Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 1–7
Marseille, 24 June 2022

NewYeS: A Corpus of New Year's Speeches with a Comparative Analysis

Anna Tramarin, Carlo Strapparava
University of Trento, FBK-IRST
anna.tramarin@studenti.unitn.it, strappa@fbk.eu

Abstract

This paper introduces the NewYeS corpus, a multilingual corpus that contains the Christmas messages and New Year's speeches held at the end of the year by the heads of state of different European countries (namely Denmark, France, Italy, Norway, Spain and the United Kingdom). The corpus was collected via web scraping of the speech transcripts available online. A comparative analysis was conducted to examine some of the cultural differences showing through the texts, namely a frequency distribution analysis of the term *"God"* and the identification of the three most frequent content words per year, with a focus on years in which significant historical events happened. An analysis of positive and negative emotion scores, examined along with the frequency of religious references, was carried out for those countries whose languages are supported by LIWC, a tool for sentiment analysis. The corpus is available for further analyses, both comparative (across countries) and diachronic (over the years).

Keywords: Political Speeches, Multilingual Corpus, Text Analysis, Computational Social Science

1. Introduction

In many European countries, it is traditional for the head of state to hold a speech towards the end of the year, which may take place on Christmas Day or on New Year's Eve. The tradition generally started in the second post-war period - with exceptions such as the United Kingdom, where it started in 1932 (Catsiapis, 2001). The speeches were first broadcast on radio and later on television, and the tradition still continues today.

It is mainly a moment of reflection on the events of the year that is about to end, which is concluded with the head of state usually expressing a positive message for the upcoming year. It is a ritual event in a country's political life, not made to convince or persuade the audience, but to create a sense of community and national identity in a moment of passage. At the same time, it serves as an implicit legitimacy of the institutions and political system that the head of state represents (Tuzzi, 2008).

Within the framework of a ritual ceremony, the personality and subjectiveness of the speaker still emerges in the choice of tone, topics and vocabulary, which is particularly noticeable in the case of elected representatives - e.g., the Italian and French presidents, who are elected every seven and five years, respectively (formerly seven in France until the year 2000) - compared to monarchs' lifetime positions. Nevertheless, the institutional tone and cultural framework is maintained across representatives, while diachronic change is perceivable in the chronological evolution of style and vocabulary (Leblanc, 2016).

In this work, we introduce the NewYeS corpus, a collection of transcripts of Christmas messages and New Year's speeches from various European countries, that cover different periods of time starting between 1946 and 1960 (depending on transcript availability online) until 2020.

This paper is organised as follows: section 2 describes the formal structure of this particular kind of speech, drawing from relevant literature; section 3 illustrates how the dataset was built; section 4 explains the analysis conducted on the texts. In section 5 we present and discuss the results, whereas in section 6 we outline possible future directions.

Country	Speech held by
Denmark	King/Queen of Denmark
France	President of the French Republic
Italy	President of the Italian Republic
Norway	King of Norway
Spain	Francisco Franco (until 1974) King of Spain
United Kingdom	Queen of the United Kingdom

Table 1: List of countries and country representatives holding the speech at the end of the year.

2. Background

The final speech of the year is usually held by the king or queen of the country - in case the country's form of government is a constitutional monarchy - or by the elected president in case of a democracy (see Table 1 for more details about the countries considered in this paper), with the notable exception of dictator Francisco Franco who held the Christmas Message in Spain until 1974.

Analysis of political discourse has been drawing increasing attention in the field of Natural Language Processing (NLP). Some work has been carried out with

regard to political stance detection (Diermeier et al., 2012; Zirn, 2014; Glavaš et al., 2017; Lehmann and Derczynski, 2019), analysis of persuasiveness in political speeches (Guerini et al., 2008), political affiliation (Navarretta and Hansen, 2020), sentiment analysis (Onyimadu et al., 2013; Abercrombie and Batista-Navarro, 2018) and topic classification in political manifestos (Zirn et al., 2016).

However, New Year's speeches represent a specific type of "institutional speech", as they are a ritual of seasonal passage between a "before" and an "after" (Tuzzi, 2008). The monarch or president also act as a symbol of the institutions, and therefore contribute to legitimise the political system they represent in front of the entire nation. Their speech emphasises a particular notion of national identity - i.e., the power's representation of national identity, addressed to an audience that is believed - and lead to - share that sense of belonging (Madsen, 2017). It aims at building the "We" in a moment of passage and potential change, hence carrier of inherent uncertainty (Van Gennep, 2013). At the same time, it tries to make sense of past events and to point out the priorities for the coming year (Leblanc, 2016). As a political ritual, the speeches present some text characteristics that may be similar across countries. For instance, a possibly fixed opening line that mainly depends on the person pronouncing the speech - e.g., the Norwegian *"Kjære landsmenn"* ("Dear countrymen"), Francisco Franco's *"Españoles!"* ("Spanish people!") or the greetings of some Italian presidents *"Italiani"* ("Italians"), *"Care Italiane, cari Italiani"* ("Dear Italian women, dear Italian men"). Sometimes even a simple "good evening" can open the speech (Spain and Italy). All the speeches usually end with the speaker wishing a merry Christmas or a happy New Year, and possibly invoking God's protection (Denmark).

The nation building effort also goes through an appeal to national values - or the values the nation is supposed to identify itself with - and an overview of the year's events as a common and shared experience.

A comparative analysis of these speeches may thus reveal both cultural differences and possible similarities, providing some insight into a tradition that for decades has been carried on in parallel in several European countries.

3. Dataset

The NewYeS corpus comprises transcripts of Christmas messages and New Year's speeches held by the heads of state - presidents or monarchs - of six different European countries (Italy, France, Spain, United Kingdom, Denmark and Norway), spanning from varying starting years to 2020. The Danish corpus is the largest, as it begins with the 1946 speech, whereas the French one is the smallest, since transcripts were available only starting from 1960.

The corpus was collected from various web sources, mainly constituted by - but not limited to - the official royal or presidential websites. In some cases the scraped texts needed a thorough cleaning before being used for NLP analysis. A primary condition for the dataset creation has therefore been the availability of the speech transcripts online.

4. Analysis

The analysis was carried out on 366 speeches (61 per country), spanning a time period of sixty years - from 1960 until 2020. This time span was chosen because speeches covering this period were available for each country, thus allowing a uniform analysis. However, some of the collected corpora contain speeches dating back to before 1960.

A first level of analysis consisted of examining the absolute frequency (i.e., the absolute number of occurrences for each year) of the word "God" throughout the texts, as shown in figure 1.

We further considered the three most frequent content words for each year - i.e., nouns, verbs (excluding auxiliary verbs), adjectives and adverbs. In particular, the word lemmas were considered for this analysis. Adverbs with a more grammatical role in the sentence, that could therefore be considered as function words (e.g., *so, too, yet, then*) were excluded. An essential step for this analysis was part-of-speech (POS) tagging, which was performed using Stanza[1], an NLP toolkit developed by Stanford University (Qi et al., 2020). An example of the three most frequent words for the year 1989 can be seen in table 2, whereas a more comprehensive comparison between years is illustrated in table 4.

Year	Country	Most frequent words (lemmas)
1989	Denmark	år (year) **frihed (freedom)** ny (new)
	France	**liberté (freedom)** pays (country) est (East)
	Italy	**libertà (freedom)** popolo (people) nuovo (new)
	Norway	mange (many) år (year) stor (big)
	Spain	desear (to wish) buen (good) hacer (to do/to make)
	UK	child world make

Table 2: Examples of the top three most frequent content words (nouns, adjectives, verbs or adverbs) as lemmas in the speeches of the year 1989.

[1] https://stanfordnlp.github.io/stanza/

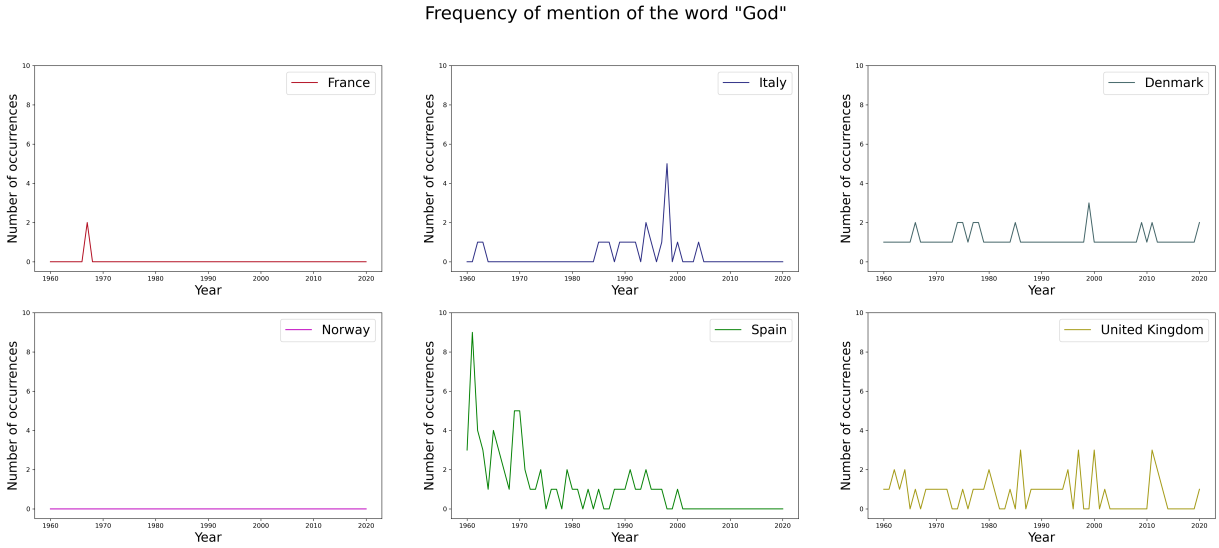

Figure 1: Absolute frequency of mention of the word "God" across countries over the years. It can be observed how the green line, representing Spain, spikes in the early '60s and slightly peaks again in the early '70s, when Franco was in power, then drops to a level that is comparable to other countries when the king of Spain started holding the speech. On the other hand, the word "God" was never mentioned in Norwegian speeches and barely in French ones (the only mention happened when some lines by the French poet Paul Verlaine were quoted, in 1967).

In addition, we applied Linguistic Inquiry and Word Count 2015 (LIWC2015)[2] for speech sentiment analysis. LIWC2015 is an application that relies on internal dictionaries in which words have been classified in different categories, according to psychological theories and emotion-measuring scales (Pennebaker et al., 2015). The tool processes the text sequentially, comparing each encountered word to the words in the dictionary and assigning scores depending on the category (or categories) a word is assigned to. The result is a speech psychometric analysis with percentage scores for each conceptual dimension. LIWC has been used in the past to identify texts written by people with mental disorders (Coppersmith et al., 2014), to analyse the correlation between narcissism and language (Holtzman et al., 2019), in social psychology (Klauke et al., 2020) and political science studies (Bond et al., 2017), among others.

The speech sentiment analysis was carried out for English, French, Italian and Spanish, as LIWC is available for these languages and has been deemed a valuable tool for multilingual analysis (Dudău and Sava, 2021). In particular, positive and negative emotion scores for specific years were extracted and compared (see Table 3), together with religion-related words. Figure 2 shows the change of positive emotion rate over time, whereas figure 3 presents the frequency of religion-related words.

[2]The LIWC2015 website: http://liwc.wpengine.com/

5. Discussion

The first analysis regarded the absolute frequency of mention of the word *"God"* in the texts. Figure 1 shows the absolute frequency of occurrence of the word *"God"* over the years. It stands out how in Spanish Christmas messages, God was mentioned significantly more often in Francisco Franco's speeches (until 1974, the year of his last speech), compared to the speeches pronounced later by the kings of Spain. A peak can also be observed in Italian speeches in 1998, when President Oscar Luigi Scalfaro held his last New Year's speech, whereas mentioning God dropped with the following Italian presidents. An interesting case is represented by Norway, as the kings never mentioned the word *"God"*. The Norwegian case is followed closely by France, since the only mention of God in 1967 was due to Charles De Gaulle quoting a line by the poet Paul Verlaine.

With regard to the most frequent content words per year, table 2 shows that *"freedom"* was one of the most repeated words in 1989, likely related to the events of that year.

A more comprehensive comparison, which takes into account different years that are representative of the whole period, can be seen in table 4. Perhaps unsurprisingly, *"year"* is one of the most repeated words overall. It is also interesting to see how *"all"* (which taken as a lemma, could stand for "every", "everything" or "everybody") seems to be a recurring word in Danish speeches, and also makes it to the top three in one French speech. It is most likely expression of an attempt to address and embrace the whole nation, in or-

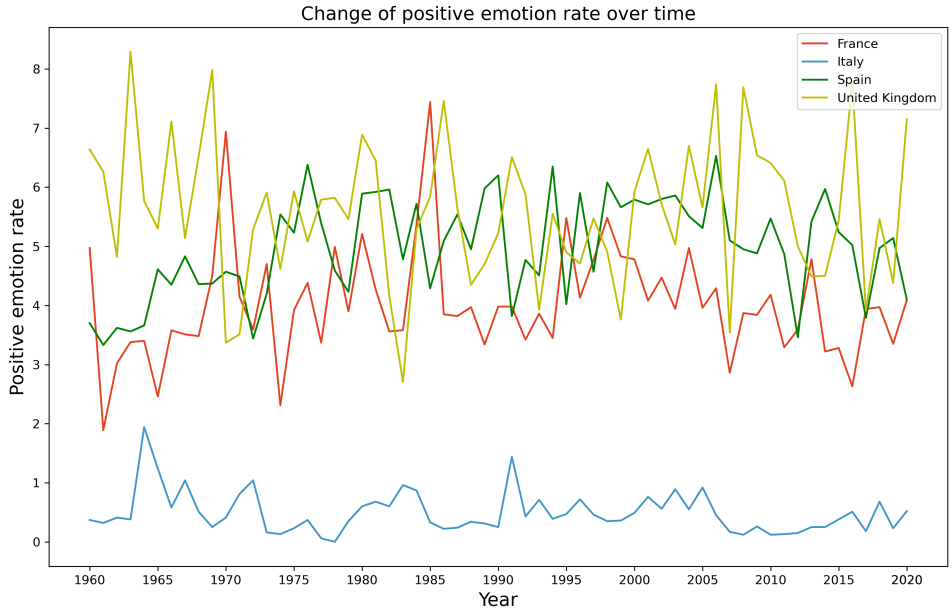

Figure 2: Change of positive emotion rate over the years

der to build a sense of community.

In addition, it can be observed how the adjective *"French"* is a recurring word in earlier French speeches, whereas the shift from *"Italian"* in 1980 to *"European"* in 2020 with respect to Italian speeches is certainly a sign of changing times.

On the other hand, some word occurrences are simply related to peculiarities of that year's speech, such as the repetition of the verb *"ring"* in the UK speech of 1980, which is due to the Queen quoting a poem by Alfred Tennyson[3].

Figure 2 gives an overview of how the positive emotion rate changed over the years, whereas table 3 shows positive and negative emotion scores assigned to New Year's speeches held in pivotal years of recent history, compared across countries. It is interesting to observe how Italy always scores very low with respect to positive emotion, regardless of the president, whereas other countries - in particular the United Kingdom - score fairly high on the positive emotion scale. This may be due to a tendency of Italian Presidents to talk about problems and challenges encountered throughout the year, whereas other speakers seem to lean more towards optimism and hope for the future.

The diachronic frequency of religion-related words is featured in figure 3. It can be seen that France presents the lowest score, followed by Italy. In line with the frequency of mention of the word *"God"*, Spain presents a higher rate of religious references when the Christmas speeches were held by Francisco Franco. Perhaps unsurprisingly, since the Queen of the United Kingdom is also head of the Church of England, her Christmas messages tend to mix political and religious elements. This distinctive trait may also contribute to the peaks in positive emotion scores, as a speech held on Christmas Day would likely be characterised by positive feelings

[3]Some verses from the poem "In Memoriam (Ring out, wild bells)"

Year	**France**		**Italy**		**Spain**		**United Kingdom**	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
1968	3.48	2.39	**0.51**	1.28	4.36	1.52	**6.52**	1.78
1973	4.70	1.46	**0.16**	2.56	4.20	2.18	**5.91**	1.43
1989	3.34	0.92	**0.31**	2.09	**5.98**	1.84	4.70	1.86
2000	4.78	1.66	**0.49**	1.03	5.79	1.42	**5.92**	1.15
2008	3.87	3.06	**0.12**	2.21	4.95	1.93	**7.69**	1.16
2020	4.09	1.64	**0.52**	1.22	4.10	1.80	**7.15**	1.16

Table 3: Positive and negative emotion scores per country in different years, in which significant historical events happened (i.e., protests of 1968, oil crisis of 1973, fall of the Berlin wall in 1989, turn of the millennium in the year 2000, financial crisis of 2008, COVID-19 pandemic in 2020).

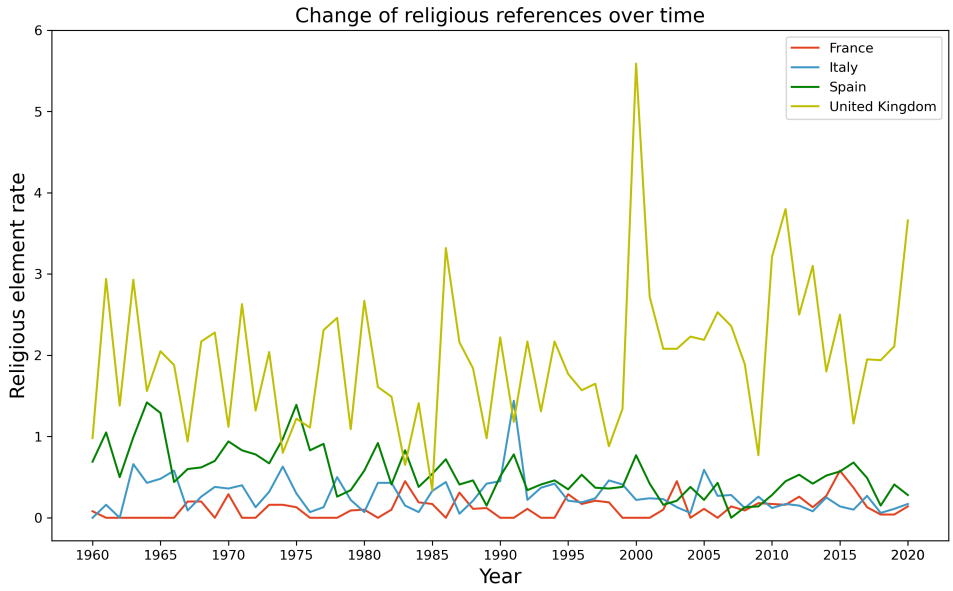

Figure 3: Frequency of religion-related words over the years

and good wishes. The highest peak of religious references can be observed in the message of the year 2000, which features the highest number of occurrences of words such as *"Christ"*, *"Christian"* and *"faith"*.

6. Conclusion

In this paper, we collected and analysed the texts of Christmas messages and New Year's speeches from six European countries. For the analysis we selected speeches that covered the same time span, namely 61 years from 1960 to 2020, for a total amount of 366 speeches. We examined the absolute frequency of mention of the word *"God"* and extracted the three most frequent content words in the speeches of specific years, which revealed some remarkable cultural differences among countries. We further implemented the LIWC2015 application for the analysis of positive and negative emotion scores for four countries (France, Italy, Spain and the United Kingdom).

This work has been an exploratory study of a subgenre of political speeches that differ considerably from the usual political talks, whose main goal is generally to convince and persuade the audience to share the speaker's opinion. New Year's speeches are a more "institutional" kind of speech held by the head of state, whose role is to represent a country's political institutions in front of the nation in a ritual way. The analysis that was implemented is an experimental example of how this corpus could be used in the framework of political analysis in NLP.

The NewYeS corpus can certainly be expanded to other countries that present a Christmas or New Year's speech tradition. For instance, at the moment transcripts of German speeches are available only from the late Eighties until today, but they would constitute a valuable addition.

With regard to future directions, focusing on one country would allow for a deeper diachronic analysis - e.g., how the vocabulary, syntax complexity and way of addressing the nation have changed over time. From a cross-cultural perspective, a detailed analysis of discourse and rhetorical strategies could highlight further differences or similarities among countries. The NewYeS corpus is publicly available for research purposes upon request to the authors.

7. Bibliographical References

Abercrombie, G. and Batista-Navarro, R. (2018). A sentiment-labelled corpus of hansard parliamentary debate speeches. *Proceedings of ParlaCLARIN. Common Language Resources and Technology Infrastructure (CLARIN)*.

Bond, G. D., Holman, R. D., Eggert, J.-A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W., Ceniceros, E. C., and Rustige, R. (2017). 'Lyin'Ted', 'Crooked Hillary', and 'Deceptive Donald': Language of lies in the 2016 US presidential debates. *Applied Cognitive Psychology*, 31(6):668–677.

Catsiapis, H. (2001). The Queen's Christmas messages. In *Seeing things: Literature and the visual. Papers from the Fifth International British Council Symposium*, pages 73–88.

Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Diermeier, D., Godbout, J.-F., Yu, B., and Kaufmann,

Year	Country	Most frequent words (lemmas)	Year	Country	Most frequent words (lemmas)
1960	Denmark	**år (year)** **al (all)** stor (big)	1980	Denmark	**år (year)** tid (time) land (country)
	France	**français (French)** bien (good/well) **tout (all)**		France	**année (year)** pays (country) **français (French)**
	Italy	problema (problem) popolo (people) **anno (year)**		Italy	**italiano (Italian)** popolo (people) giovane (young)
	Norway	**år (year)** menneske (man/human being) god (good)		Norway	**år (year)** dag (day) bli (to become)
	Spain	político (political) **año (year)** social (social)		Spain	esfuerzo (effort) querer (to want) mejor (better)
	UK	time good **year**		UK	service ring (Verb) many
2000	Denmark	**al (all)** familie (family) god (good)	2020	Denmark	mange (many) **år (year)** **al (all)**
	France	**année (year)** nouveau (new) avoir (to have)		France	**année (year)** avoir (to have) vie (life)
	Italy	**anno (year)** avere (to have) fare (to do)		Italy	**anno (year)** paese (country) **europeo (European)**
	Norway	stor (big) **år (year)** tid (time)		Norway	**år (year)** bli (to become) god (good)
	Spain	**año (year)** hoy (today) noche (evening/night)		Spain	gran (big) sociedad (society) tener (to have/to have to)
	UK	**year** life man		UK	light hope **year**

Table 4: Examples of the top three most frequent content words (nouns, adjectives, verbs and adverbs) as lemmas from the speeches of four different years, with the corresponding translation for languages other than English.

S. (2012). Language and ideology in Congress. *British Journal of Political Science*, 42(1):31–55.

Dudău, D. P. and Sava, F. A. (2021). Performing multilingual analysis with Linguistic Inquiry and Word Count 2015 (liwc2015). an equivalence study of four languages. *Frontiers in Psychology*, 12:2860.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2017). Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693.

Guerini, M., Strapparava, C., and Stock, O. (2008). Trusting politicians' words (for persuasive NLP). In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 263–274. Springer.

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., et al. (2019). Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples. *Journal of Language and Social Psychology*, 38(5-6):773–786.

Klauke, F., Müller-Frommeyer, L. C., and Kauffeld, S. (2020). Writing about the silence: identifying the language of ostracism. *Journal of Language and Social Psychology*, 39(5-6):751–763.

Leblanc, J.-M. (2016). *Analyses lexicométriques des vœux présidentiels*. ISTE Group.

Lehmann, R. and Derczynski, L. (2019). Political Stance in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages

197–207.

Madsen, C. (2017). Magtens repræsentation af danskhed. konstitueringen af nationalidentitet i Dronning Margrethes nytårstaler. *Passage-Tidsskrift for litteratur og kritik*, 31(76).

Navarretta, C. and Hansen, D. H. (2020). Identifying parties in manifestos and parliament speeches. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 51–57.

Onyimadu, O., Nakata, K., Wilson, T., Macken, D., and Liu, K. (2013). Towards sentiment analysis on parliamentary debates in hansard. In *Joint international semantic technology conference*, pages 48–50. Springer.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Tuzzi, A. (2008). Messaggi dal Colle: l'analisi statistica dei dati testuali e il discorso di fine anno del Presidente della Repubblica Italiana. *Messaggi dal Colle*, pages 1000–1021.

Van Gennep, A. (2013). *The rites of passage*. Routledge.

Zirn, C., Glavaš, G., Nanni, F., Eichorts, J., and Stuckenschmidt, H. (2016). *Classifying topics and detecting topic shifts in political manifestos*. University of Zagreb.

Zirn, C. (2014). Analyzing positions and topics in political discussions of the German Bundestag. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 26–33.