The Role of Language in the Automatic Coding of Political Texts

Didier Ruedin

didier.ruedin@wolfson.oxon.org

POST-PRINT

This is the final draft *after* refereeing

Published as: Ruedin, Ruedin, Didier. 2013. "The Role of Language in the Automatic Coding of Political Texts." *Swiss Political Science Review* 19 (4): 539–45. doi:doi:10.1111/spsr.12050.

http://onlinelibrary.wiley.com/doi/10.1111/spsr.12050/full

Replication data: http://hdl.handle.net/1902.1/20 302

Abstract

Automatic approaches to coding party manifestos and other political texts have become more widespread. This research note addresses the question to what extent the source language of a text affects the results. To do so, Swiss manifestos in German and French are coded automatically, comparing a keyword-based dictionary approach and Wordscores. Because of language differences, both stemming and particularly stop words are important to obtain comparable results for Wordscores. If both are used, the predicted scores are almost identical in both languages. With the right preparations, the challenge of language differences can thus be overcome.

The Role of Language in the Automatic Coding of Political Texts

3 February 2013

Abstract

Automatic approaches to coding party manifestos and other political texts have become more widespread. This research note addresses the question to what extent the source language of a text affects the results. To do so, Swiss manifestos in German and French are coded automatically, comparing a keyword-based dictionary approach and Wordscores. Because of language differences, both stemming and particularly stop words are important to obtain comparable results for Wordscores. If both are used, the predicted scores are almost identical in both languages. With the right preparations, the challenge of language differences can thus be overcome.

Introduction

There has been significant progress on automatic coding of political texts in recent years, and such approaches are becoming more commonplace (e.g. Laver and Garry, 2000; Laver et al., 2003; Grimmer, 2010; Grimmer and King, 2011). Possibly the most common application in the political sciences is the automatic coding of party manifestos to obtain party positions, but there are also other applications. This research note will use automatic coding of party manifestos as an example to illustrate the way differences in language can affect empirical results. In comparative research language differences are largely a confounding factor when data are derived from texts. Such differences affect surveys where translation effects may occur (Behling and Law, 2000; Davidov and De Beuckelaer, 2010), but also automatic approaches where words without substantive content – such as articles and prepositions – may influence results.

A number of methods have been developed for coding party manifestos automatically. There are two key attractions to using automatic approaches for coding party manifestos. First, compared to manual approaches computerized methods are resource friendly and perfectly reliable. Second, compared to expert surveys, text-based approaches can easily be expanded backward over time to address questions not considered at the time when expert surveys were carried out. Particularly in comparative research, the availability of automatic coding has opened up new perspectives. This is due to the unprecedented amount of data that can be compared, and the ease with which some of the automatic approaches can be adapted to different contexts. Since correlational validity has been demonstrated for most automatic approaches (e.g. Klemmensen et al., 2007), automatic coding is now a largely accepted method.

This said, there are unresolved questions, such as the influence of words without substantive contents in comparative contexts. It is important to ensure that automatic approaches work independent of the source language. For example, it would be unacceptable if say the positions obtained from French manifestos were systematically biased to the left compared to British manifestos. In normal circumstances it is difficult to separate the influence of language from the difference of national contexts, given that the two tend to vary jointly. This means, that the assumption that the source language has no impact on the estimated positions largely

remains untested. Some confidence can be gained from the successful application of automatic approaches in a cross-national setting bridging language differences (e.g. Klemmensen et al., 2007).

It is clear that automatic methods such as Wordscores work for many languages, in the sense that predicted party positions correlate highly with expert positions and manual coding of party manifestos. The question remains, however, to what extent reported results are independent of the source language. Given that the same party manifesto is not normally available in multiple languages, a direct test is difficult. On the one hand, some of the automatic approaches use differences in languages, notably in the sense that parties on the political left and parties on the right tend to use different words and phrases in debates about the same political issue. On the other hand, from a methodological point of view, there are undesired differences in languages in terms of grammar and vocabulary in use, fixed expressions that may exist in some languages, or the use of similes and metaphors (Crystal, 2007).

There has not been much research on the impact of the source language on results from automatic approaches. To a limited extent studies that compare methods can be useful, in the sense that not all methods are affected by differences in the source language to the same extent (e.g. Chen, 2011; Klemmensen et al., 2007; Laver et al., 2006; Debus, 2009). For example, Klemmensen et al. (2007) compared different methods across Western languages. They demonstrated that automatic methods seem to work across languages, but there is no direct comparison of the same text across languages. Similarly, Giger et al. (2011) applied Wordscores to the multilingual setting of Switzerland, but they ran two parallel analyses: one for the German-speaking cantons, and one for the French-speaking cantons. In terms of understanding the influence of the source language on estimated party positions, this is comparable to Klemmensen et al.

A direct comparison of the same manifestos in different languages was undertaken by Collette and Pétry (2010) who examined the effect of language on Wordscores and Wordfish estimates in Canada. They compared three parties between 2000 and 2008 and report high correlations between languages, especially for the stemmed Wordfish estimates. They do not seem to take into consideration that correlation coefficients can be expected to be high in their case because of similar party positions over time. They found that stemming manifestos increases the correlation between estimates from different languages for Wordfish, although no results for word stemming in the case of Wordscores are reported. The use of stemming is in line with the recommendation by Slapin and Proksch (2008) to prepare texts accordingly (see also Lowe, 2008).

Data and Methods

To examine the impact of the source language on the results from automatic approaches, this research note uses the party manifestos from 13 Swiss parties of the 2011 election. In Switzerland, party manifestos are faithful translations, and this was manually verified for large sections of the manifestos in question. By faithful translations I refer to the fact that the manifestos are translated sentence by sentence – Collette and Pétry (2010) use the expression ‘parallel texts’. Despite this, differences between source languages can be expected because the exact expressions used are not always the same, after all, manifestos use natural language. Covered are parties across the political spectrum, including all major parties as well as some smaller ones: AL, BDP, CSP, CVP, EDU, FDP, GPS, JEVP, JGPS, Pirates, SD, SPS, and SVP. For the

EDU and the JEVP a mission statement of similar character to the party manifestos was used, but the reported results in no way depend on the inclusion of these two cases.

Two automatic methods are used in this research note: Wordscores and a key-word based dictionary. Throughout the research note, the interest is on the predicted party positions using the different methods. Wordscores use word frequencies in the manifestos and reference texts with 'known position' to identify party positions (Lowe, 2008). I use Will Lowe's Wordscores implementation Austin for R (Lowe, 2011; R Development Core Team, 2012). Where stemming is used, this was done in JFreq (Lowe, 2010). The 20 most common words in the manifestos of each language were designated stop words.[1] Wordscores and the dictionary approach were used to estimate positions in a specific policy domain (immigration), although these positions are very highly correlated with left-right positions. Concerns about language should be unaffected by the choice of policy domain. As computerized methods, both Wordscores and the dictionary approach are inherently agnostic as to what issue domain is analysed. The GPS and the SVP were used as reference texts in Wordscores, set to the positions in Ladner et al. (2009). Substantively the same results can be achieved with different reference texts. I have checked the estimated party positions against expert data (Benoit and Laver, 2006; Ladner et al., 2009) to verify face validity.

For the dictionary approach, a dictionary of keywords was developed and tested extensively across countries and languages – including languages not covered in this research note. Translation and back-translation from and to English were used to ensure a high degree of equivalence of the dictionaries in different languages (Behling and Law, 2000). The dictionary was used in conjunction with Will Lowe's Yoshikoder (Lowe, 2009), which reports both the count of matches, and a rate taking into account the length of the manifestos. For the dictionary approach stemming makes no difference, since the keywords used match to stems. Similarly, stop words make no real difference, because these common words are not matched by the keywords. As such, the reported counts are the same. For the reported rates, stop words could potentially make a substantive difference, if the stop words have very different frequencies in the different languages.

Findings

First, let us look at the results for plain manifestos, that is manifestos as they are. The estimated party positions for the two source languages correlate highly, but there remain noticeable differences. The number of keyword matches is similar for both languages, but not identical for all the manifestos (r=0.99). By contrast, the correlation is lower for the rate of matches – taking into consideration the length of the manifesto – at 0.97. These differences between languages are not entirely unexpected, and we probably simply look at differences in the frequency of the most common words in each language. The count of matches is not affected by the overall length of the manifesto, hence a higher correlation coefficient.

[1] Lowe and Benoit (2013 forthcoming) highlight that stop words could carry information, even if they appear à priori uninformative. The results presented here suggest that 20 stop words are an adequate number, but this may not be the definitive answer.

For the Wordscores estimates, the correlation between the two languages is 0.88 when using plain manifestos. This points to differences in word frequencies in the two languages, despite the faithful translations. The results are in line with Collette and Pétry (2010), although because of the different samples involved we cannot directly compare the correlation coefficients. What is more, Collette and Pétry examine party positions over time and pool estimates when calculating correlations.

As a second step, the words in the manifestos were stemmed. For example, the words *immigrants*, and *immigration* are both reduced to *imigr*. For the dictionary approach, stemming makes no difference, since the keywords were designed to match stems anyway. For Wordscores, we observe trivial differences. For the specification used in this research note, we observe a small reduction in the correlation coefficient from 0.88 to 0.81. Most commonly there are no differences, or a small increase, such as in a specification using subsections of the manifestos: r=0.74 for the plain manifestos, and r=0.80 for the stemmed manifestos. Despite trying many different specifications, the differences never were significant – usually there were insignificant increases.

As a third step, both stemming and stop words were applied. Whereas stemming reduces words to their roots, stop words indicate a list of words that are removed from analysis. In this case, I used the 20 most common words for each language as stop words. This means that some of the words in the list are not equivalents in the two languages. Once we apply both modifications, for Wordscores the correlation between languages increases to 0.995, a significant increase compared to the 0.88 for plain manifestos. In fact, the correlation between languages is near perfect once both stemming and stop words are applied.

For comparison, Wordscores applied to a subsection of manifesto, as was done above, lead to a correlation between languages of 0.97. This is also a nearly perfect correlation, in line with other specifications. Moreover, these results reflect findings by Collette and Pétry (2010) who found marked increases in correlation coefficients – albeit for the Wordfish method.

Table 1 summarizes the correlations between the two languages for five different methods. For the dictionary approach, we find very high correlations for the rate of matches, but particularly for the count of keyword matches. For Wordscores, the estimates using plain manifestos correlate highly, but are clearly not identical. Using a stemmer to reduce words to their roots does not have the desired effect. However, when applying both a stemmer and stop words, the correlation coefficients indicate a near perfect relationship between the estimates derived from either source language.

	Dict (rate)	Dict (count)	Word (plain)	Word (stem)	Word (stop)
r=	0.97	0.99	0.88	0.81	1.00

Table 1: Correlations between estimates based on German and French estimates respectively. The reported correlations are between the estimated party positions for 13 parties in both languages. The methods are a keyword-based dictionary approach ('Dict'), both as a rate and a pure count, as well as Wordscores ('Word') on plain manifestos, stemmed manifestos ('stem'), and with both stemming and stop words applied ('stop').

Discussion and Conclusion

The departing question of this research note was whether the source language of political texts affects results that use these texts as data. I used party manifestos as a

means to derive party positions to illustrate that the influence of language on results can be significant. Swiss party manifestos were used, because they offer faithfully translated manifestos. This means that the influence of language could be isolated – apart from the language the compared manifestos are identical –; in comparative research, by contrast, linguistic and other differences are conflated.

The results indicate that if nothing is done and we use plain manifestos, language differences can indeed affect the results of automatic approaches. For keyword matches the differences are small, but the reported rate of matches is not exactly the same for each language. For Wordscores the differences between languages are significant, unless both a stemmer and stop words are applied. The reported differences resonate findings for Canada, where French and English manifestos were compared (Collette and Pétry, 2010). In contrast to Collette and Pétry, however, I use a significantly larger number of parties, and report findings for Wordscores. I find that stemming manifestos is insufficient to compensate for language differences. In this case, the source language can clearly have an impact on the estimated positions – even where manifestos are translated sentence by sentence. It appears that using stop words is a necessary intervention to obtain comparable results: a nearly perfect correlation between party positions in the two languages. Indeed, it is the stop words, not the stemming that seems to make the big difference. This said, even à priori uninformative words can carry information for placing parties (Lowe and Benoit, 2013 forthcoming), so the list of stop words should probably be limited in length. In this research note, using the 20 most common words in each language as stop words led to satisfactory results.

The keyword-based dictionary approach was not much affected by differences in language, but for analyses over time, the changing vocabulary of political debates can pose a validity problem. More generally, a dictionary approach relies on (in-depth) knowledge of the researcher, which can be a challenge in comparative studies involving many countries – and with that, many languages. Fully computerized methods such as Wordscores are more flexible in this regard, since new reference texts can be added as the debate evolves.

The results in this research note are likely to hold for other political texts and automatic approaches to assigning values. They also have direct bearing on comparative research where texts across languages are included, and usually separate analyses are used for each language. In this context, differences in language can confound results. However, with the right preparations – if both stemming and stop words are used – it seems that differences in language are no longer a problem for automatic approaches: The estimates are nearly identical, irrespective of the language of the manifesto used. Put differently, the challenge of differences in language can be overcome with the right preparations: word stemming and particularly the use of stop words.

References

Behling, O. and K. Law (2000). *Translating Questionnaires and Other Research Instruments: Problems and Solutions*. London: Sage.

Benoit, K. and M. Laver (2006). *Party Policy in Modern Democracies*. London: Routledge.

Chen, Y. (2011). Quantitative content analysis of Chinese texts?: A methodological note. *Journal of Chinese Political Science* 16, 431–43.

Collette, B. and F. Pétry (2010). Comparing the location of Canadian political parties using French and English manifestos as textual data. Paper presented at the *From Text to Political Positions Mining for Meaning* conference, Amsterdam. http://www2.let.vu.nl/oz/cltl/t2pp/docs/ws2010/papers/P3-Collette.pdf

Crystal, D. (2007). *How Language Works: How Babies Babble, Words Change Meaning and Languages Live or Die*. London: Penguin.

Davidov, E. and A. De Beuckelaer (2010). How harmful are survey translations? A test with Schwartz's human values instrument. *International Journal of Public Opinion Research* 22 (4), 485 –510.

Debus, M. (2009). Pre-electoral commitments and government formation. *Public Choice* 138 (1), 45–64.

Giger, N., J. Müller, and M. Debus (2011). Die Bedeutung des regionalen Kontexts für die programmatische Positionierung von Schweizer Kantonalparteien. *Swiss Political Science Review* 17 (3), 259–85.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis* 18 (1), 1–35.

Grimmer, J. and G. King (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences* 108 (7), 2643 –2650.

Klemmensen, R., S. Hobolt, and M. Hansen (2007). Estimating policy positions using political texts: An evaluation of the Wordscores approach. *Electoral Studies* 26 (4), 746–755.

Ladner, A., D. Schwarz, and J. Fivaz (2009). Die Schweizer Parteien im politischen Raum - eine Analyse der politischen Positionen ihrer Kandidierenden bei den Nationalratswahlen 2007. *Working Paper de l'IDHEAP* 2009 (1).

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review* 97 (02), 311–331.

Laver, M., K. Benoit, and N. Sauger (2006). Policy competition in the 2002 French legislative and presidential elections. *European Journal of Political Research* 45 (4), 667–697.

Laver, M. and J. Garry (2000). Estimating policy positions from political texts. *American Journal of Political Science* 44 (3), 619–634.

Lowe, W. (2008). Understanding Wordscores. *Political Analysis* 16 (4), 356.

Lowe, W. (2009). Yoshikoder: Cross-platform multilingual content analysis. Java software version 0.6.4, http://www.yoshikoder.org

Lowe, W. (2010). JFreq: Count words, quickly. Java software version 0.2.5, http://www.conjugateprior.org/software/jfreq/

Lowe, W. (2011). Austin: Do things with words. R package version 0.2, https://r-forge.r-project.org/projects/austin/

Lowe, W., and K. Benoit (2013 forthcoming). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Slapin, J. and S. Proksch (2008). A scaling model for estimating timeseries party positions from texts. *American Journal of Political Science* 52 (3), 705–722.