Policy regimes, latent constitutive institutions, and machine learning methodology – examples from EU energy policy and City of Helsinki environmental policy

Arho Toikka

Paper to be presented at the Vincent and Elinor Ostrom Workshop in Political Theory and Policy Analysis colloquium series, 12th November 2014

Abstract: Besides laying out formal rules-in-use, policy documents (legislation, planning documents, strategies, roadmaps) also define the constitutive settings of policy regimes – i.e. what do we talk about when we talk about energy policy? What counts as an energy policy argument? This paper uses a big data approach and a machine learning methodology, topic modelling, to map the latent structure in two energy and environmental policy corpuses and interprets them through an institutionalist framework. Topic modelling finds the statistical structure of words appearing together in documents and forming "topics". If these topics are to be meaningful for social sciences, they need to have theoretically informed interpretability. Previous research has interpreted them as frames, discourses, and semantic fields. This paper explores the structure of European Union energy policy and City of Helsinki environmental policy to put forward the argument that they could be interpreted as classifying or constitutive institutions.

This paper is draft version is written with upcoming funding applications in mind, while practically working with the methods, keeping an eye on eventually publishing an analysis, using a preliminary data set. As such, it is structured like a research paper, but includes some things disproportionately (musings on data handling) while glossing over others (discussion and implications).

Introduction

The energy regime - or the constellation of networks of actors, material artefacts, and rules associated with them, governing how energy is supplied, converted, and used – is slow to change. This paper argues that 1) (some of) this inertia comes from constitutive rules or cultural-cognitive institutions that define the regime, used in energy regime texts, and 2) that these rules are implicit, based in mental models held by actors, potentially unknown to the text authors themselves, 3) the relationships between these rules are key to understanding the system, and 4) that machine learning can be engaged to find this structure.

Classifications and categorizations are important in human interactions. Sometimes they are simple and explicit – i.e. a person born at least 18 years ago is an adult. Sometimes they are complex and implicit – i.e. whether the use of carbon capture to use the CO_2 captured in a marketable product qualifies as emission reductions or trading credits within an emission trading scheme, a debate currently being settled in the EU. In sociological institutionalism, these common frameworks of meaning are central to the concept of institutions (Scott 2001). This research has focused on large-scale embeddedness of institutions in their historical context on the one hand and on the habitualization of everyday practices on the other (Gronow 2008), while other institutionalist traditions have focused on those rules and rule-like phenomena evident in language, even if implicitly and without punishment mechanisms or normative consequences (Crawford & Ostrom 1995). In this paper, these two traditions converge: classifications and cultural scripts are important, but they are specific, lingual institutions used in specific decision-making settings to make sense of the complex underlying phenomena. They are the constitutive definitions policy actors use (but also amend and change) when arguing on public policy.

Just as rules of football constitute the possibility of playing the game (Searle 1969), the possibility of writing an energy policy document only arises when actors hold shared definitions of what constitutes the energy regime. But unlike football, these rules are not spelled out, they are implicit. What counts as 'social sustainability' does not have an easy definition, but rather it is constituted by multiple different statements, dispersed throughout the institutional text. And these definitions form systems or networks: when talking about, say, a coal-fired power plant, an author implicitly talks about a set of technologies and social practices within the physical power station and outside it, in mining practice, electricity transmission and so forth. When exogenous pressures, social networks and technological solutions change, what is implied and constituted may not change in parallel. New socio-technical solutions to new issues may require a different alignment of constitutive elements.

The goal of this paper is to apply a machine learning methodology called topic modelling to two different energy regime data sets or corpuses and explore a constitutive institutional interpretation of the results. Topic models are unsupervised models that find a latent, unobserved structure in observed text by modelling word co-occurrence. The idea is that each text is about a small number of topics, each topic defined as probability distribution over the observed vocabulary, and the goal of the analysis is to find the unobserved topics and distributions that would have been most likely to generate the observed texts. Whether the reorganization of data into a topic structure is a meaningful thing and not just a statistical artefact requires a measure of content validity and an associated theoretical framework, as there is no goodness-of-fit statistic for an unsupervised method. Here, the resulting topics and word probabilities are interpreted as defining institutions, and the validity of the models is measured in how much sense they make.

The computational or automated coding and analysis of meaning in text data has emerged as a useful technique for many social science research questions in the recent decades thanks to an explosion of computing power and easily available computer-readable text (Ventresca & Mohr 2002; Grimmer & Stewart 2013 provides a recent review). The methods do not compete with each other, but different methods are useful for different research questions and different data sets (Grimmer & Stewart 2013). There are roughly two different approaches to machine learning from text: supervised and unsupervised methods. Supervised methods and many early analyses relied on human-generated word lists or training data (human-coded examples from where the computer learns the coding scheme) to do analysis. These supervised methods are interested in discrete classification and the ability of computers to replicate human placement of texts into classes (Purpura & Hillard 2006) or measuring a quantified, theoretically defined variable from text, like placement of author on the left-right scale (Slapin & Proksch 2008). Now methods that are able to find structure unknown to the analysts beforehand are emerging. These unsupervised methods can have different goals, but usually the aim is to classify the documents in the data set into either discrete classes or mixed probabilities.

Most commonly, they still model for a single clustering, but mixed-membership models have also been used, topic modelling forming at least some part of the analysis in these cases (Grimmer & Stewart 2013). A recent special issue in *Poetics* (Mohr & Bogdanov 2013) demonstrated the use of topic modelling on a range of social scientific text corpuses, including newspaper articles, news stories, U.S. National Security Strategy Reports, scientific papers and works of fiction. The topics were interpreted as frames, scenes, or broad themes. In political science, such examples include Grimmer (2010) who uses topic modelling to find expressed political agendas in senate press releases, and Quinn et al (2010) who use topic modelling to find the main topic of each speech in the U.S. Senate Congressional Record. In these analyses, each document is still placed in a single category or topic, and mixed-membership is a measure of senators publishing press releases on multiple topics (Grimmer 2010) or dynamically over time, with different days having multiple topics discussed (Quinn et al. 2010).

The approach taken here differs from the existing literature in the nature of texts analyzed and the assumptions of the relationships between those texts and the institutions expressed in those texts. First, instead of speeches or press-releases by policy actors, this paper uses longer public policy texts. The texts are specifically chosen so that they define the wide-ranging issues area. Thus, the methodology will have to acknowledge that each document consists of multiple documents. In fact, the relative proportions of topics and the similarity of these proportions between these topics is at the heart of the analysis. Accordingly, the method chosen retains the information about the mixed-membership of the documents.

This paper uses two data sets to demonstrate this idea, one from the European Union and the other from City of Helsinki metropolitan policy. The number of documents in the data sets is small (for a machine learning analysis), but the documents included can be rather extensive; meaning any hand-coding attempt to quantify them would be a considerable task.

The paper is organized as follows: The next section describes the two case studies. After that, the methodology and practical application of topic modelling is presented, followed by a presentation of the institutional framework suggested for the interpretation of the topic model. Then, the structure of the two case studies is analyzed.

Background and data

This paper uses two cross-sectional data sets to explore the theoretical and methodological challenges at hand. These are meant to represent two parts of the nested or multilevel governance research on decentralization of energy regimes in Finland. Currently, the two cases are the EU and the City of Helsinki. The selection of cases was due to organizational constraints: this analysis is meant to be the first step in the analysis of institutional constraints to decentralization of energy regime in Finland, and they build on earlier projects. In the future, these will be complemented with a Finnish national data set and rural case studies, as well as longitudinal aspect. The data sets are not directly comparable.

Defining the population of relevant documents is not a straight-forward task. The goal is to map the multi-actor rule system that interacts to make a semi-coherent policy regime. There is no single actor with power to make decisions over the regime, but even those with formal authority refer to other organizations and actors, formally devolving some of their powers or informally trusting the expertise of others. Whereas, say, the output of a senator easily be defined as the speeches he gives, European energy policy is not solely defined in the directives of the European Commission. A variety of political and bureaucratic organizations is involved in making the decisions. Here, a simple quick-and-dirty procedure was used: the collection of the documents was chosen by the authorities themselves. A few key communication channels where both legislation and strategy documents are posted were identified, and this selection was used as-is.

The first, the EU energy regime data, consists of documents that were published by the European Commission Directorate-Generals on Energy and Climate Action on their websites. This is arguably a reasonably complete picture of an EU energy regime – the DG-Clima is a spinoff from the larger DG-Environment, and energy issues with an environmental focus were moved into this new unit. This data set holds 46 documents, listed in Appendix A.

The second data set represents the environmental policy of the City of Helsinki and the metropolitan area around it, as represented by the municipal collaboration organization Helsinki Region Environmental Services. The environmental policy implementation and practical level work is largely organized by the City Environment Centre and decisions are taken or approved by either the City Council or the City Board. The reason for the more general environmental focus instead of a strict energy focus comes from the organization of responsibilities in Finland and the resulting organizational structure: the city does not formally have an energy policy, but it owns a large-energy company run as fully-owned subsidiary of the city, and governing over the actions of the city company and all others is done via various environmental policy instruments, including zoning practices and such. Still, documents strictly laying out methods of environmental protection were excluded. The analysis was run in Finnish and the final results translated to English. The data set contains 40 documents published by these organizations.

From a text corpus to a topic structure

Topic modeling (Blei 2012) is a method for finding the latent or underlying structure of a text corpus. The intuition behind topic models is to think of documents as if they were generated by the author choosing a topic, picking a word according to the distribution of all words within that topic, and placing it in the document. The task of the modeler is to reverse-engineer the intents of the author by finding what topic distribution is most likely to have generated the observed documents. (Mohr & Bogdanov 2013.) The results are probabilities of words used within a topic, and distribution of topics in each document.

The topic model used here is the simplest version, called Latent Dirichlet Allocation (LDA; Blei 2002). LDA is multi-level Bayesian model, inferring the posterior distribution of topics, words in topics, topic of each word, and the topic mixtures in documents on the basis of the observed words in documents. This inference problem is intractable and various algorithms exist to approximate it. In this paper, the variational Expectation Maximation method was used.

Obviously, sampling from topics is not how text in real life is generated and topic modelling does not aim to reveal the mechanism of policy-making. Thus, whether the results of the analysis have meaningful social science implications depends on if the technical assumptions to allow modelling can be justified in terms of the research questions, and if the resulting topic structure has a meaningful theoretical justification. The technical key assumption of the models is to define meaning as wholly relational, so syntax, rhetoric and grammar and inconsequential. This translates to the so-called bag-of-words assumption, meaning that the method ignores word order, grammar, and syntax.

Topic modelling is an unsupervised method, meaning that there are no human-coded correct answers to a partial data set to learn from. This means that there is no natural to analyze the goodness-of-fit within the method or the data. Resampling methodology can estimate the stability of the results, but internal validity has to be judged by an external reference, either evaluation by human experts, semantic experiments (Chang et al. 2009), or by evaluation the meaningfulness of the results by reference to an external framework, such as the theoretical institutional approach adopted here.

The analysis presented in this paper is draft version. Topic modelling requires a fair amount of data manipulation, and this manipulation is not final, but rather should be taken as a proof-of-concept. This is due to using fairly ad hoc selections for a) the number of terms retained for the analysis, and b) the number of topics interpreted.

Practically, the analysis requires the preprocessing of the corpus, the creation of a numerical representation of the data as a document-term matrix, running the analysis and evaluating the results with the help of substantive theory, and communicating the results. The analyses was done using the programming language R and packages available for R, especially the tm package (Feinerer 2014) and the topicmodels package (Grun & Hornik 2011). These packages offer tools for preprocessing the raw documents into proper bags-of-words; that is, with certain stopwords (function words, prepositions etc.) removed as per a stopword dictionary, words stemmed, punctuation removed, and so forth, and tools for translating them into document-term frequencies. In the analyses here, the document-term matrices were further pruned by using the term-frequency inverse-document frequency statistic. Term frequency-inverse document frequency counts how often a word occurs in a given document, but weighs that by how often it appears in the whole corpus. The goal was to remove words too common to be interesting, like commission, Helsinki, and so forth. A better, domain specific stopword dictionary could be used instead. English language stemming is quite straight-forward, but the nature of Finnish with the heavy use of inflexion to communicate person, case, etc. makes it more complicated. The current analysis uses the standard Snowbell stemmer with R tm implementation, though better alternatives might be available (Kettunen & Baskaya 2011).

Regardless of these challenges, the two corpuses were successfully transformed into document-term matrices. For the Helsinki analysis, the resulting matrix has 1726 unique terms, forming 15 topics. For the EU analysis, there are 1309 terms, forming 20 topics. Grimmer (2010, 12) and Quinn et al. (2019) use substantive interpretation to decide the number of topics, and a similar procedure is followed here. These choices were made by varying the cut-off in term-frequency inverse-document frequency and the number of topics with the goal of minimizing single-topic documents and single-document topics as the upper-limit. Heuristically, these were not substantially interesting, coupling words from a document title with a few terms used in the particular document most often. At the lower limit of topics, terminologies started fusing together in ways that appeared to combine two issues, often based on one document known to handle those two. At the lower limit of terminology, overlap between documents disappeared.

Topics and institutions

What is the relationship between "X counts as Y in context C" and a distribution of words in a body of policy text? I am claiming they can be treated as if they were the same thing, and this is not a straight-forward claim. The claim treats rules as linguistic entities, but not once that exist statements. The claim implies that rules are not only shared, but they are laid out collectively by a set of actors not necessary in touch with each other. The claim states that regulation has institutional implications beyond the formal and these implications combine with those put forward in informal institutional text. The following section discusses whether these claims can be justified in terms of institutional traditions.

Ostrom (2005, 19) defined working rules as those "to which participants would make reference if asked to explain and justify their actions to fellow participants", but they may need the help of an institutional analyst digging deep to reveal these rules to them. The topic model digs out word distributions, but these do not correspond to the lingual form of strategies, norms, and rules with deontics. They are complex representations of the attributes and qualifiers that drive system understanding.

Cultural-cognitive institutions are paradigms or mental models that strongly influence system understanding, boundary setting, and search-space for solution space (Pahl-Wostl 2009). The constitutive schema is legitimated not by reference to sanctions or moral acceptance, but by how comprehensible and recognizable it is within the relevant culture (Scott 2001). Superficially, these accounts of deeply internalized or habitual institutions seem to contradict the importance of language. The method proposed here is a way to bridge this apparent gap by enabling the analysis of the lingual form of these deeper institutions. Not all habits are institutional: many are simply straight-forward solutions to acting in the physical world and apparent shared understanding is just rational actors reaching the same conclusion. Tacit knowledge held by a community is real phenomenon, but it not necessary to relax the requirement that institutions are linguistic (or symbolic).

In a public policy setting, the problem is building on earlier knowledge to find an institutional solution that fits the technological and social settings. Institutional structure is critical to the degree (and process) of integration of diverse knowledge (North 2005). Like any social order, the energy regime is a problem solution system that communicates knowledge and skills from a variety of scientific disciplines and social practices to enable decision-making (Aligica & Boettke 2009). In any reasonably complex system, some of this communication will be between the lines: it is not feasible to write the communication as an algorithmic rulebook defining the processes. Any argument or attempt to bring new knowledge into the institutionalized system builds on other institutionalized forms of knowledge, amending and changing them. Social learning is a distributed process that cannot be represented a sum individual learning (Pahl-Wostl et al., 2007).

Then, the institutional generative account of a policy text used here argues that whatever agenda an author is trying to communicate, they draw on a set of established constitutive rules. To add to the collective body of knowledge, the author of a new text uses information in individual rules and information about the system of rules, or the interdependencies of rules. These rules are directly unobservable and not actively known to the users. The rules form a nested search space for authors to use and change; in this view policy argument consists of taking the established rule-set and amending it. This suggests change is usually incremental and potentially path dependent (Pierson 2000), as modern policy often is. But whether something is incremental or revolutionary also depends on the viewpoint of the analyst, and might be just as much a function of the scope of analysis as the scale of change.

In a dynamic topic model, the incrementality of change also becomes an empirical question: if change mostly happens with topics, amending the established vocabulary, change is incremental; if there is more radical change, the proportions of topics used should change.

Currently, the analysis treats authorship and actors as irrelevant, as if there was a single truly shared rule system. This is inaccurate. Of course, there are often multiple authors for a given text, occupying different roles in the process, and different authors form networks of shared meaning. This can potentially be incorporated into the method with extensions that allow for use of metadata.

Analysis

The analysis here has two parts: first, the interpretation of topics and then an exploration of the relationships between topics (and documents). In future work, validity and sensitivity will need to be attended to – now I will save a few words about how that might proceed and what problems there are in the discussion. The analyses presented here explore two slightly different ways to explore the topic structure and its implications. For the EU analysis, the analysis draws on social network analysis and presents a networked based on the similarity of documents (in terms of shared topics), as well as the similarity of topics (in terms of shared documents). For the Finnish analysis, the presentation focuses on the interpretation of technical or issue-area topics versus the structural institutions of time-scale, measurement, and sustainability.

Topic interpretation and labeling

For the European Union analysis, table 1 displays the top ten most common words within a selection of the observed topics, in their stemmed form. They were selected to demonstrate the three categories the topics fall into: intuitively, the topics discussed in the corpus seems to focus on either a) technologies and policy instruments, b) the implementation, organization, and monitoring of policy, or methodological issues and c) wider policy issues, potentially having to do with implications and such. The terminology in the topics varies from the technical (table in modelling, or shall in biofuels, shall being heavily used in EU Directives) to the substantial terminology. The nature of EU text is also reflected in the heavy presence of abbreviations.

Future economy	Asian relations	Biofuels	Forestry and agriculture	Clean development mechanism	Ghg reporting	Modelling
target	price	shall	soil	res	tier	target
stakehold	electr	articl	agricultur	scenario	tabl	total
roadmap	ruec	biofuel	eccp	baselin	inventori	model
billion	wind	electr	target	target	nir	assum
job	japan	target	sequestr	nsat	ipcc	refer
lowcarbon	period	refer	coeq	price	total	electr
agricultur	retail	bioliquid	forest	memberst	plant	consumpt
skill	china	consumpt	monitor	ghg	submiss	price
period	consumpt	ethanol	gase	prime	ghg	plant
electr	shale	paragraph	electr	cdm	crf	scenario

Table 1. Selection of topics from EU corpus topic mode

The fact that these multiple dimensions arise in the data is central to the institutional argument put forward here. Usually, political science topic models have focused on finding the underlying collection of issues, corresponding to the organizational setting of policy making, for example using the match between topics and legislative committees as validation of the model (Quinn et al 2010). It is apparent that these topics are not single policy issues.

Each topic represents the established terminology to discuss an issue. The topics are constituted by the documents, some literally: Clean Development Mechanism of the Kyoto protocol and its European implementation is actually defined in a few of the documents included in the analysis. The methodological topics draw on terminology and references where to documents, policies, and treaties are defined, such as nir (National Inventory Reports) or the tiers defining minimal allowable reporting methods under topic greenhouse gas reporting. The methodological topics tend to be more diffuse in the corpus: 15 documents draw at least 5% of their words from modelling , but no document draws more than 14%.

The technical topics relate to fuel resources, conversion methods, emission reduction methods, or energy use and efficiency. The more strictly energy topics are named biofuels, bioenergy, electricity, coal & natural gas, shale gas, carbon capture & storage, and nuclear power. Dealing with emission reductions through other instruments is reflected on in topics of forestry & agriculture, land use issues and land use scenarios. These topics are typically more centered in few documents: emission scenarios for non-CO2 emitting energy by 2030 are 80% of words from the bioenergy topic, reflecting the centrality of making and implementing national biomass plans around the time these documents were written in 2008. On the other, topics such as shale gas only made up 67% of the European Commission communication on fracking and shale gas – the communication commented also on renewables, carbon capture & storage, and land-use.

Perhaps the most interesting feature of the topic structure is the set topics connecting energy policy to wider issues, whether having to do with international relations in general or Asia in particular or with the future economy. Before the analysis, it would have been hard to name these as themes central to European energy policy documents. They feature in the most documents, usually making up a small portion of the words on the order of 5-10%.

In the City of Helsinki analysis, the main themes are similar, but slightly different. Table 2 again has the ten most common words within the topic, but the Finnish stems have been translated to full English words. There are again three types of topics: a) those referring to issues areas of importance, b) those referring to time-scale of policy and measurement of it, and c) sustainability topics.

Energy efficiency	Social sustainability	Midterm plans	Transport	Large city plans
Megawatthours	Unemployment	Resource needs	Urban traffic	Six cities
Savings potential	Responsible organization	Implementing organization	Climate strategy	Helsinki Regional Environmental Authority
Electricity	Climate strategy	Regional zoning plan	Hybrid	Average
Savings instrument	Tourism	Blueprint management	Biodiesel	Carbon neutral
Environmental investment	Rail traffic	Climate strategy	Biking	Time period covered
Environmental cost	Nature guide	Environment centre	Electric car	Statistics Finland
Heating	Homes	Land use legislation	Charging station	Analysis weight
Energy review	Participation	Report on needs	Energy companies	Climate action

Table 2. Selection of topics from the Helsinki corpus

The issue areas in Helsinki environmental policy are energy efficiency, energy production, transport fuels, transport planning, water quality and waste-water treatment. The high salience of water issues is probably part artefact of data selection, part a reflection of the real processes: Helsinki is a coastal city with recreationally and ecologically important rivers and streams within urban areas, so it is natural that they will remain on the agenda, and there were a few notable policy processes that had to with water and other issues ongoing at the time the data was extracted. The main port of Helsinki was moved from the traditional location in what is now downtown to the outskirts, balancing ecological values with easier freight access, and opening up debate on what should be done with the coal-fired power plants close to the downtown areas that still get coal shipments.

Helsinki policy appears to be prepared with two different time-scales, mid-term plans and long-term plans, each with their own terminology and associated policy instruments. Climate oriented policy deals with the long term and produces concepts and visions to 2050 and beyond, while more practically oriented fields such as zoning and construction deal with time-scales of 5 to 10 years. There is little overlap between these topics. Right now, there is a policy debate on whether to close the coal-firing power plant located in prime residential area or retrofit it for burning biofuels, mainly woodchips and wood pellets. There is definitely a real phenomenon of both bureaucratic representatives and political actors speaking past each other. This observation might be explainable in terms of the different time-scales they draw on.

All of this is embedded in sustainability discourse on two dimensions: social issues and more generic sustainability talk. The first covers unemployment and cultural sustainability among other issues, the second might be interpreted as attempting to define metropolitan sustainability or the way a city is organized.

Topic structure

The relationships between the various topics in the European Union corpus are presented here with the help of social network analysis. Figure 1 shows the relationship between the topics, with the number of documents consisting at least 5% of words from that topic defining the size of nodes (i.e. topics) and number of shared documents defining distance between topics, represented in two-dimensional form using the social network analysis software Gephi and the Force Atlas algorithm for dimension reduction. In this representation, the policy regime seems to be centered around 4 questions: what is the relationship of individual policy to the emission trading system? How does it affect EU international commitments? How does it affect future economy in EU? And how do we model and measure the effects of these policies?

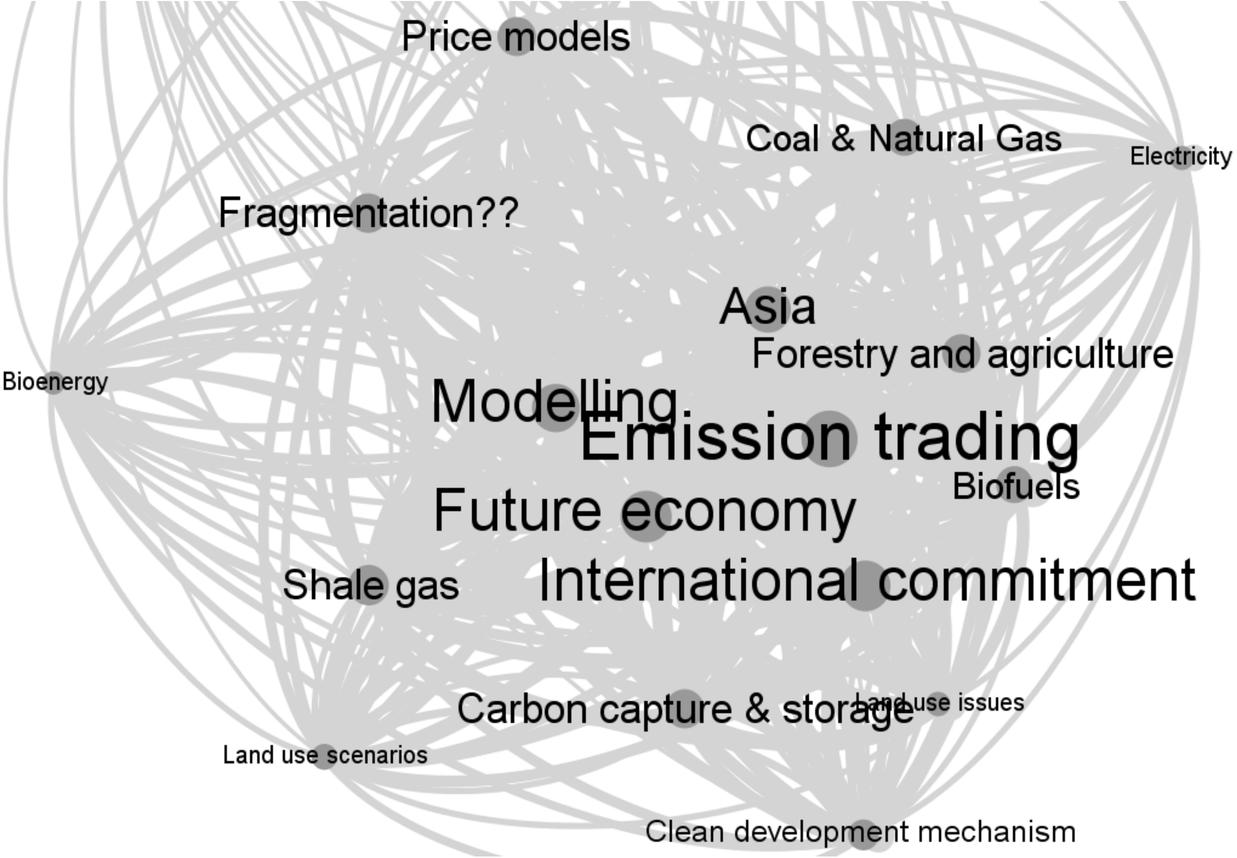


Figure 1. Relationships between topics in EU analysis

Figure 2 maps the relationships between policy documents similarly, with the size of nodes (i.e. documents) representing number of topics and distance is how many topics two documents share. The figure reveals a few interesting details: some of the most central documents are what would have been expected, like the generic 2050 roadmaps, but some of the more administrative documents that map research for priorities and options are central too. Centrality in this map does not translate to power or importance, but an analyst would probably be interested in looking at these particular documents qualitatively to see if they actually were important.

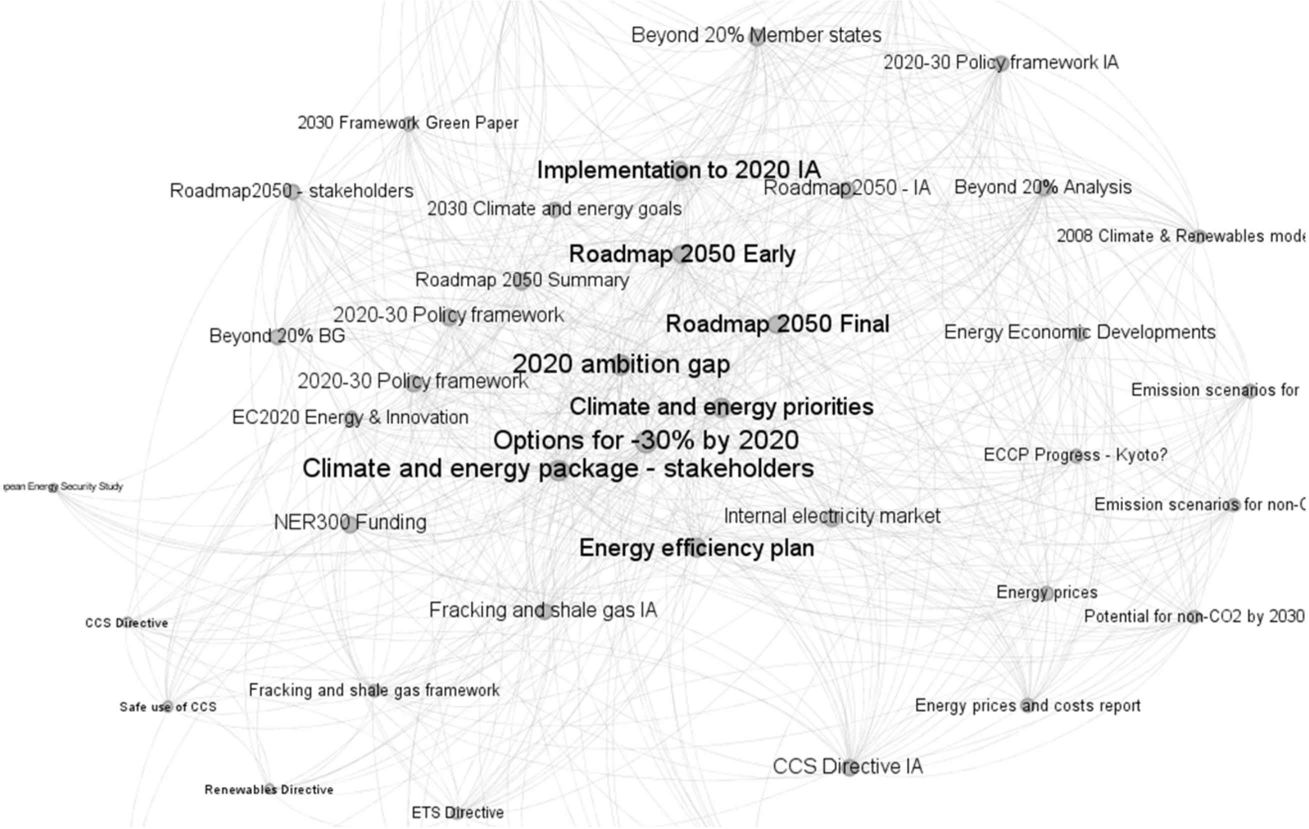

Beyond 20% Member states
2020-30 Policy framework IA
2030 Framework Green Paper
Implementation to 2020 IA
Roadmap2050 - stakeholders
2030 Climate and energy goals
Roadmap2050 - IA
Beyond 20% Analysis
2008 Climate & Renewables mode
Roadmap 2050 Early
Roadmap 2050 Summary
2020-30 Policy framework
Roadmap 2050 Final
Beyond 20% BG
Energy Economic Developments
2020 ambition gap
2020-30 Policy framework
Emission scenarios for
Climate and energy priorities
EC2020 Energy & Innovation
Options for -30% by 2020
ECCP Progress - Kyoto?
pean Energy Security Study
Climate and energy package - stakeholders
Emission scenarios for non-C
Internal electricity market
NER300 Funding
Energy efficiency plan
Energy prices
CCS Directive
Fracking and shale gas IA
Potential for non-CO2 by 2030
Fracking and shale gas framework
Energy prices and costs report
Safe use of CCS
CCS Directive IA
Renewables Directive
ETS Directive

Figure 2. Relationships between documents in EU analysis

Figure 3 is a similar map on blocks of topics grouped by the intuitive typology presented above, and the relationships describe sum of percentage of words in shared documents, i.e. if a document was 50% from one topic and 50% from another, this would be coded as 1, 10% from one and 15 from another would be from 0.25, normalized over the number of documents, focusing on sustainability. In other words, over all words drawn from the sustainability topic, 24% of the words in the same documents discuss energy, 11% water, and so forth.

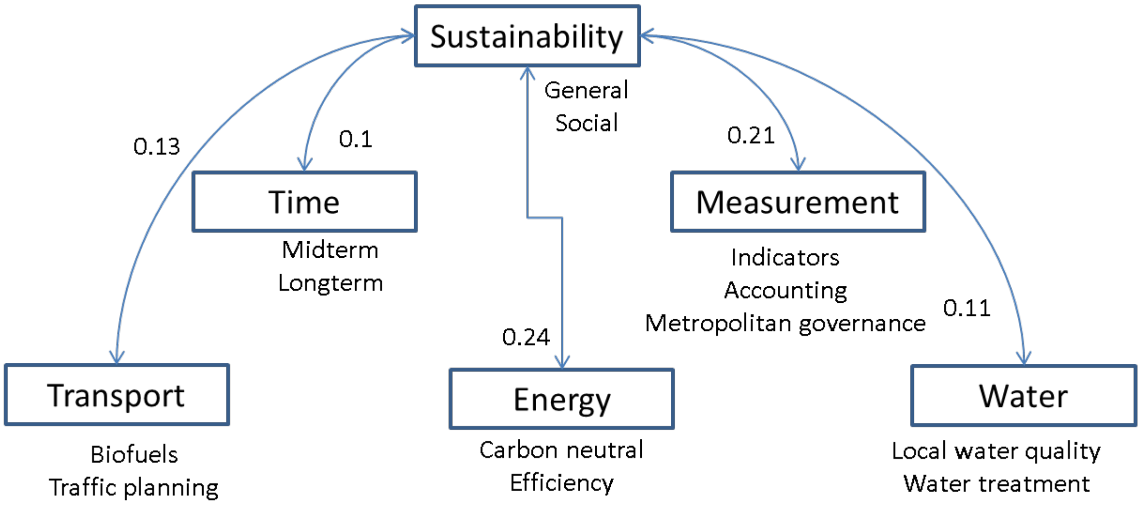


Figure 3. Relationships between sustainability topics and other topic areas in Helsinki analysis

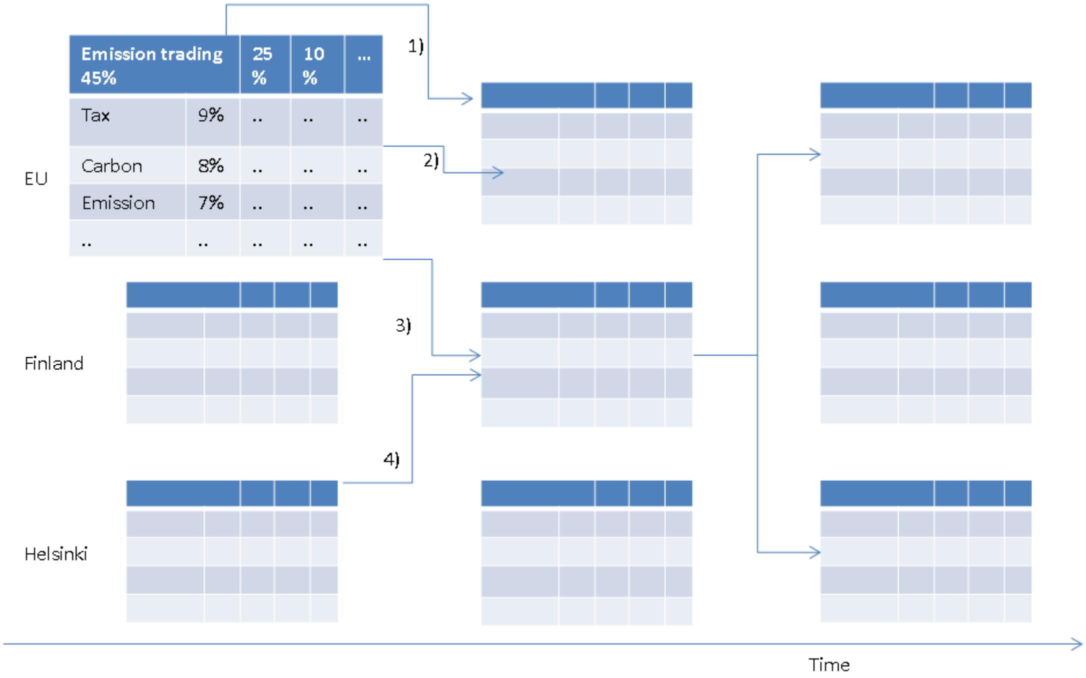
Discussion: Steps forward

The goal of this paper is to quantify or at least make measurable the intuition that how things, technologies, groups, and other entities are defined and classified in policy-making is extremely influential on final policy choice. The analyses above represent a first exploration of applying topic modelling to a public policy corpus. The goal, for now, was to demonstrate how topics in such corpuses probably do not reflect a single dimension, like summarizing distinct policy issues, but rather summarize over combinations of policy issues and instruments, their implementation, and how they link to wider societal issues. A number of issues remain to be solved to make this idea into a feasible research project with valid methodology. These include the question of how to test the validity and sensitivity of the analyses, how to draw boundaries to define the data set, and what the method for exploring between topic relationships should be.

In general, the problem of validation and is hard for an unsupervised model. Quinn et al. (2010) discuss their model in terms of semantic validity, convergent and discriminant construct validity, predictive validity and hypothesis validity. Semantically, the model here appears valid – each topic has a coherent meaning and the relationships between them are meaningful. An experimental approach as in Chang et al (2009) could easily be constructed to test this.

It's less clear what the construct validity would be. What is the competing measure to compare these to? For predictive validity, the problem is to differentiate between the influence of topics and other measures. There would need to be an external event expected to change the topic structure by a quantifiable measure, and then it could be tested if this happened.

Regardless, the next step from this analysis is to design the actual data collection process and move to a dynamic model that allows for the evolution of topics over time. The problem of boundary setting or population definition is interesting. Plausibly, all discussion on-going in the society, produced by sources from mass media to scientific debate to internal documents of companies, political parties and so forth influences the language of public policy. How to define a feasible but still comprehensive way to define which documents are meaningfully within the energy regime?



Emission trading 45%		25 %	10 %	...
Tax	9%	..	..	..
Carbon	8%	..	..	..
Emission	7%	..	..	..
..	..	..	..	..

Figure 4. Dynamic multi-level governance measured through topic modelling

Fundamentally, I am interested in using this method to uncover details on how policy text and ideas move and develop in a multi-level governance setting. Basically, the idea is measuring the topic structure of a nested energy governance setting and see how the terminology develops over time at each of the levels. Figure 4 provides an intuition of this framework. The structure can evolve in four ways: 1) the relative contribution of topics can change within a governance level, 2) the terminology used within each topic can change, or topics can influence other levels either 3) top-down or 4) bottom-up.

References

Aligica, P. D., & Boettke, P. J. (2009). Challenging institutional analysis and development: the Bloomington school. Routledge.

Blei, D.M., (2012). Probabilistic topic models. Communications of the ACM 5, 77–84.

Feinerer, I. (2014). Introduction to the tm package Text Mining in R. http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Expressed Agendas in Senate Press Releases. Political Analysis 18: 1-35.

Grimmer, J. & Stewart, B.M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis 2013: 1-31.

Gronow, A. (2008). Not by rules or choice alone: a pragmatist critique of institution theories in economics and sociology. Journal of Institutional Economics 4: 351-373.

Grun B & Hornik K (2011). topicmodels: An R Package for Fitting Topic Models."Journal of Statistical Software,40 (13), 1-30. URL: http://www.jstatsoft.org/v40/i13/

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. Neural Information Processing Systems, 2009.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. Poetics, 41(6), 545-569.

Kettunen, K., & Baskaya, F. (2011). Stemming Finnish for Information Retrieval–Comparison of an Old and a New Rule-based Stemmer. In *Proceedings of the 5th Language & Technology Conference (LTC 2011), Poznan* (pp. 476-480).
Ostrom, E. (2005). Understanding Institutional Diversity. Princeton University Press: Princeton.

Pahl-Wostl, C. (2009). A conceptual framework for analysing adaptive capacity and multi-level learning processes in resource governance regimes. Global Environmental Change, 19(3), 354-365.

Pahl-Wostl, C., Craps, M., Dewulf, A., Mostert, E., Tabara, D., & Taillieu, T. (2007). Social learning and water resources management. Ecology and Society, 12(2), 5.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. American political science review, 251-267.

Purpura, S and Hillard, D. 2006. Automated classification of congressional legislation. In Proc. Digital Government Research, pages 219–225.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209-228.

North, D. C. (2006). Understanding the process of economic change. Academic Foundation.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. American Journal of Political Science, 52(3), 705-722.

Ventresca, M.J. & Mohr, J.W. (2002). Archival Research Methods. In Baum, J.A.C (ed.): Blackwell Companion to Organizations. Oxford: Blackwell Publishers.

Appendix A - European Union energy policy documents used in the analysis

EC 2020 Energy and Innovation Conclusions
Comm. Energy prices and costs report
Comm. In-depth study of European Energy Security
To DG Env. Model-based analysis of the 2008 EU Policy Package on Climate Change and Renewables
Comm. Implementation of NER300 funding
CCS Directive Impact Assessment
Directive on Promotion of Renewable Energy
Directive on Emission Trading Scheme
Directive on Carbon Capture and Storage
Comm. Roadmap to low carbon economy 2050 Stakeholder consultation
Comm. Roadmap to low carbon economy 2050 Impact assessment
Green paper on 2030 framework for climate and energy policies
Policy framework for climate and energy 2020-2030
Comm. On fracking and shale gas
Policy framework for climate and energy 2020-2030 Impact assessment
Policy framework for climate and energy 2020-2030 Executive summary
Roadmap for moving to low carbon economy 2050 March 2011
Roadmap for moving to low carbon economy 2050 May 2011
Climate and energy priorities for Europe
EU Climate and Energy Package: Citizen's summary
Comm. Delivering the internal electricity market
Policy framework for climate and energy in the period from 2020 to 2030
Policy framework for climate and energy in the period from 2020 to 2030 Background
Policy framework for climate and energy in the period from 2020 to 2030 Impact assessment
Comm. Energy efficiency plan
Energy Economic Developments in Europe
Energy Prices in the EU Working document
Annual EU greenhouse gas inventory 90-12+14
Closing the pre-2020 'ambition gap'
Ensuring safe use of CCS in Europe
Analysis for options for reducing ghgs by 30% by 2020
2030 climate and energy goals for a competitive, secure and low carbon EU
Emission scenarios for non-CO2 gases in the EU-27
Emission scenarios for non-CO2 gases in the EU-27 Final
Potentials and costs for mitigation of non-CO2 greenhouse gases until 2030
Roadmap for moving to low carbon economy 2050 Summary
Implementation of climate change and renewable energy for 2020 Impact assessment
Options for moving beyond 20% reductions in ghgs Background
Options for moving beyond 20% reductions in ghgs Analysis
Second ECCP Programme Progress Report - Can we meet Kyoto targets?
Second ECCP Programme Progress Report - Executive summary
Analysis for options beyond 20% Member state reports
Comm. On fracking and shale gas Impact assessment
Technical report for beyond 20% reductions
EU Energy trends to 2030
EU Energy, transport and ghg emissions Trends to 2050