TOPIC MODELLING: A NOVEL METHODOLOGY FOR THE SYSTEMATIC STUDY OF HIGHER EDUCATION INTERNATIONALIZATION POLICY

Daniela Craciun, Craciun_Daniela@phd.ceu.edu, Central European University

Abstract: Higher education internationalization represents a strategic priority for governments around the world, not least because of the academic, economic, socio-cultural and political benefits expected to derive from it. Nevertheless, how policy makers understand the idea of internationalizing higher education systems differs from country to country. Because of the sheer volume of policy texts and the plethora of policy measures that come under the umbrella concept of internationalization, researchers struggle to classify and make inferences about internationalization. This state of affairs raises a salient question: how can we systematically study internationalization policies without massive costs in terms of funding and time? The paper posits that using computer assisted topic modelling techniques represents an innovative way to deal with this issue. Using an original database of national policies for the internationalization of higher education, the paper demonstrates how topics can be automatically retrieved from documents while meeting validity and reliability standards. Specifically, Latent Dirichlet Allocation (LDA) which is a generative probabilistic model is applied on the text corpora in Python so as to provide a glimpse of how public policy documents can be efficiently processed, summarized, compared and classified based on topic probabilities. In turn, this allows researchers to examine multiple cases while having limited resources and to discover new or understudied similarities between policies adopted by different countries.

Keywords: public policy, higher education, methodology, internationalization, content analysis, classification
Introduction

The internationalization of higher education emerged as a key topic in research and policy debate in the 1990s (Guruz, 2008) when the link between this process and its possible economic, political, socio-cultural and academic benefits became apparent (van der Wende, 2001). The latest *Global Survey on the Internationalization of Higher Education* reveals that internationalization “remains, or indeed grows in importance for higher education institutions” (Egron-Polak & Hudson, 2014, p. 6). Internationalization has, thus, come to be seen as the “central motor of change” (ibid, p. 5) in higher education. Consequently, policies that promote this process have developed from ad-hoc to more structural measures which aim to have a deep impact on national higher education systems (Brandenburg & de Wit, 2011; Teichler, 2009). This development has been mirrored in the scholarly arena where the internationalization of higher education has assumed greater importance as a research topic (Kreber, 2009) becoming the most studied topic in tertiary education policy (Huisman, 2008).

Nevertheless, surveying the literature on higher education internationalization reveals that “since the Second World War the concept has been understood and applied in a variable fashion” (Callan, 2000, p. 16). This failure to sort out and clarify the concept has negative implications for knowledge accumulation and cross-case comparability, hence, jeopardizing our ability to make systematic findings. With this state of affairs in mind, the present paper pleads for a more nuanced understanding of ‘internationalization’, and the policies that promote the process, so that is becomes a fact-finding category with adequate discrimination power. The paper suggests a way forward by proposing the construction of a typology of internationalization that helps to clarify the concept and against which national policy efforts can be assessed.

Indeed, heuristic tools have already proven useful in comparative research. For instance, they allow scholars to map the empirical distribution of cases and, thus, encourage rigor and enhance transparency in case selection (Elman, 2005). In higher education research, classificatory frameworks have been developed for systematizing policies related to higher education regimes, institutional internationalization, educational monitoring, international education hubs, international scholarship programs and institutional differentiation. Thus, the paper proposes that constructing a similar typology of national higher education internationalization strategies would enhance the transparency of both academic endeavors and policy practice.
The aim of the paper is not to convince the reader of the merits of a higher education internationalization typology\(^1\), but to outline a way to build this heuristic device without massive costs in terms of funding and time. Therefore, the paper proceeds as follows. The first section will shortly and purposefully review the literature on internationalization in higher education policy trying to point out the problematic issues in existing research. The second section introduces topic modelling as a computer assisted content analysis method that can help in building a typology. Specifically, Latent Dirichlet Allocation (LDA) which is a generative probabilistic model is applied on the text corpora in Python so as to provide a glimpse of how public policy documents can be efficiently processed, summarized, compared and classified based on topic probabilities. Finally, the paper will close with the expected contributions of the methodology for public policy analysis.

**Systematizing Knowledge on Internationalization**

Getting to the heart of what internationalization means is not a simple matter” (van Gyn et al., 2009, p.27). Throughout the years, definitions of internationalization have evolved in various ways: from focusing on a set of specific activities to be carried out by universities (Arum & van de Water, 1992)\(^2\) to viewing it as a dynamic process to be integrated in the wider set of organizational activities of higher education institutions (Knight, 1993)\(^3\); from focusing on internationalization as an institutional endeavor to viewing it as a result of broader developments and synergies between various levels of authority with the power to steer internationalization (van der Wende, 2001)\(^4\); and, finally, from viewing internationalization as a limited function of a university’s context (Soderqvist, 2002)\(^5\) to viewing it as a broad and eclectic mix of policies and processes that evolve on various scales (Knight, 2003b)\(^6\).

\(^1\) For an in depth analysis of this issue see Craciun (2015)

\(^2\) Internationalization represents “the multiple activities, programs and services that fall within international studies, international education exchange and technical cooperation” (Arum & van de Water, 1992, p.202)

\(^3\) Internationalization is “the process of integrating an international/intercultural dimension into the teaching, research and service functions of the institution” (Knight, 1993, p.21)

\(^4\) Internationalization is “any systematic, sustained effort at making higher education (more) responsive to the requirements and challenges related to the globalization of societies, economy and labor markets” (van der Wende, 1996, p.23)

\(^5\) Internationalization is “a change process from a national higher education institution to an international higher education institution leading to the inclusion of an international dimension in all aspects of its holistic management in order to enhance the quality of teaching and learning and to achieve the desired competencies” (Soderqvist, 2002, p.29)

\(^6\) Internationalization is “the process of integrating an international, intercultural or global dimension into the purpose, functions or delivery of post-secondary education” (Knight, 2003b, p.2)
Crucially, mastering the literature on internationalization is made difficult by:

(1) the proliferation of different labels associated with the historical development of internationalization: “re-internationalization” (Teichler, 2009), “postinternationalization” (Brandenburg & de Wit, 2011), “globalized internationalization” (Jones & de Wit, 2014);

(2) the coexistence of various scales of analysis: “internationalization versus Europeanization” (Teichler, 2002), “internationalization versus globalization” (Scott, 2000), “Europeanization versus internationalization versus globalization” (Callan, 2000);

(3) the development of similar parallel processes across different scales: “de-monopolization”, “de-institutionalization”, “de-nationalization” (Kehm, 2003); and

(4) the plethora of policy measures that come under the ‘banner concept of internationalization’ (Callan, 2000), i.e. international student mobility, collaborative research, development of curricula and strategies for teaching and learning, the establishment of institutional networks (Altbach, Reisbeg & Rumbley, 2009).

Another reason, and maybe the most important one for this lack of conceptual clarity, is the perpetual quest for generalization in the social sciences. The ubiquitous use of the concept of internationalization (Teichler, 2009) has resulted in what is known “as the Hegelian night in which all the cows look black and eventually the milkman is taken for a cow” (Satori, 1970, p.64). In other words, trying to obtain worldwide applicability has led to concept stretching which, in turn, has reduced the analytical purchase (Dunlop & Radaelli, 2013) of internationalization. This is problematic, not only for academic research, but also for public and institutional policy formulation, because it suggests that scholars talk past each other, making the accumulation of knowledge increasingly difficult (ibid.).

If the trend continues and “internationalization is underscored as an argument for almost any higher education reform” (Teichler, 2009, p.94), the results will invariably be disappointing. Thus, the paper argues that internationalization policies should not be sold as a panacea for all the problems of higher education. Instead, it should be acknowledged that internationalization policy for higher education is no one-size-fits all approach. Rather, it is an umbrella term that comprises various measures designed to tackle specific issues (Callan, 2000). How policy makers understand the idea of internationalizing higher education systems differs from country to country (Graf, 2009; King,
2010; Matei & Iwinska, 2014) which goes to prove that internationalization is far from being a one-dimensional concept (Kreber, 2009).

All in all, it is safe to argue that there is no consensus on the scope of meaning of internationalization in higher education. Considering the importance attached to this process by universities (Egron-Polak & Hudson, 2014; European University Association, 2013), governments (Jones & de Wit, 2014; Kalvemark & van der Wende, 1997; Lujiten-Lub, van der Wende & Huisman, 2005; Knight, 2004), supranational organizations and institutions (European Commission, 2013; Henard, Diamond & Roseveare, 2012), non-governmental actors (British Council, 2011), and scholars (Huisman, 2008), it is puzzling to understand why there is not more agreement on the conceptual borders of internationalization. This state of affair raises a salient question: how can the conceptual clarity of internationalization be improved so as to increase its explanatory purchase in the study of higher education? The paper posits that developing a typology of internationalization is a viable solution to this conundrum. However, the question of how we can we systematically study internationalization policies without massive costs in terms of funding and time remains. The paper posits that using computer assisted topic modelling techniques represents an innovative way to deal with this issue.

Research Design and Data Collection

In order to answer the research question and build the proposed typology, I suggest a research design that makes use of computer assisted content analysis to examine national policies for the internationalization of higher education. The main objective is to establish similarities and differences between the morphological constructions of policies for internationalization and thus, develop a better understanding of the concept. In what follows, I will detail what the components of a computer assisted content analysis research design are, how the primary data for the analysis was collected, and how the method proposed works for analyzing the data.

Components of a Content Analysis Research Design

As defined by Krippendorff, a founding father of this scientific method, “content analysis is any research technique for making replicative and valid inferences from data to their context” (1980, p.21). Hence, content analysis can be applied so as to objectively quantify the existence of certain
words, concepts, themes, sentences, phrases, idioms, or characters in texts, establishing a link between their content and their institutional, social and cultural context (Berg, 2001). I will return to this issue later on. For now, it is enough to say that this method is particularly useful for tackling the question proposed by the paper as it can reveal priorities of internationalization policies.

The physical makeup of any content analysis research design are: unitizing, sampling, recording/coding, reducing data, abductively inferring, and narrating the answer to the research question (Krippendorff, 2004). Next, these constitutive parts will be discussed in more detail.

Firstly, unitizing refers to “the systematic distinguishing of segments of text (…) that are of interest to an analysis” (ibid., p.83). In other words, it means making decisions about the units that will inform the content analysis exercise. In turn, units are entities that the researcher differentiates and considers as independent (ibid.). At this point it is important to distinguish between two types of units: sampling unit and coding unit.

On the one hand, sampling units represent the entities that are “distinguished for selective inclusion in an analysis” (ibid., p.98), i.e. the texts to be analyzed (newspaper articles, policy documents, images, etc.). In the context of my research, the national plans affecting the internationalization of tertiary education will serve as the sampling unit for the study. The list of documents includes both stand alone national internationalization strategies, and general national higher education policies that include references to internationalization.

On the other hand, coding units represent the entities that are “distinguished for separate description, transcription, recording, or coding” (ibid. p.98), i.e. the units that are to be categorized (paragraph, phrase, word, etc.). In the context of the paper, as in most other research using computer assisted content analysis, the word will represent the unit of coding. I will return to this issue soon when the issue of recording/coding data is discussed.

---

7 Except for the last two components – abductively inferring and narrating the answer to the research question – which are currently work in progress in my PhD thesis and will be added to the discussion once the data analysis phase is finished.

8 Internationalization plans “can include goal statements, mission statements, vision statements, implementation initiatives, allocated resources, timelines and performance indicators” (Childress, 2009, p.3) that express a country’s vision of internationalization.
Secondly, sampling in a content analysis research design refers to the same process like in survey research – but instead of sampling people from a population, the researcher samples texts from a corpora, which is a collection of populations of text. In the context of my research, the corpora of texts refers to all the policy documents directed at internationalizing higher education. Due to a number of theoretical reasons, the corpus (the population of texts to be analyzed) of my research represents a census of national strategies for the internationalization of post-secondary education. The rationale behind conducting a census is that there is no reliable population list of higher education internationalization policies from which documents could be selected using a reliable sampling techniques.

Before moving on to the next constitutive component of the research design, it seems necessary to explain why the proposed analysis will be carried out at the national policy level. Understanding the role that the nation-state plays in the internationalization of higher education is a crucial issue both in the academic literature and in policy practice. The distinction between empirical evidence and normative issues is not always clear cut. Nevertheless, it is hardly disputed that development of modern higher education institutions is closely linked to the state:

“Thereir regulatory and funding context was, and still is, national; their contribution to national cultures was, and still is, significant; students tended to be, and still are, trained to become national functionaries; and universities played, and still play, a considerable role in what some have called the military-industrial complex of the nation state.” (Enders, 2004, p.365)

In fact, research has shown that national policies and the national context play the most significant role in the internationalization of higher education (Enders, 2004; Luijten-Lub, van der Wende & Huisman 2005; Graf, 2009). Moreover, both higher education institutions (European University Association, 2013; Egron-Polak & Hudson, 2014) and supranational organizations (European Commission, 2013; Henard, Diamond & Roseveare, 2012) expect and encourage the participation of the state in this process. As a result, there are calls for more centralized and broader strategic approaches\textsuperscript{9} to internationalization and harmonization of policies across sectors (i.e. labor market,

\textsuperscript{9} Comprehensive strategies are believed to include: international mobility of students, scholars and administrative staff; integration of an international dimension in the curricula; international research collaboration; establishment of institutional networks and programs; and cross-border delivery of education.
migration, trade, economic development, foreign affairs) so as to address both national and institutional interests (Enders, 2004; British Council, 2011; European Commission, 2013).

Thus, the decision to conduct the analysis at the national level was taken for a number of reasons. Firstly, as a plethora of studies have shown, nation states still play a central role when it comes to steering HE (Beerkens, 2004; Enders, 2004; Vlk, 2006; Witte, 2006). As such, higher policy “still tends not only to reflect but to underscore the specific traditions and circumstances of individual countries” (Enders, 2004, p. 361). Secondly, these plans express a political commitment to internationalization, and not just a political rhetoric. In other words, they can be considered part and parcel of the policy output of any government that promotes a supportive culture towards internationalization. Also, such plans push governments to operationalize their understanding of internationalization. Having a well-defined and coherent national strategy has been shown to be an important ingredient for moving forward with internationalization efforts (British Council, 2011; Henard, Diamond & Roseveare, 2012). Lastly, the advantage of employing this strategy is that the unit of analysis remains constant on a cross-national basis. Moreover, it helps to establish the parameters of the study and represents a guide for data sourcing (Yin, 2003).

Thirdly, recording/coding “bridges the gap between unitized texts and someone’s reading of them, between distinct images and what people see in them, or between separate observations and their situational interpretations” (Krippendorff, 2004, p.84). In the context of my research, this task will be carried out using automatic computer coding. Such an analysis assumes that texts are bags of words. This means that the order of the words in the text is not important for the analysis and thus, it is discarded. As such, it is assumed that “a simple list of words (…) is often sufficient to convey the general meaning of a text” (Grimmer & Stewart, 2013, p.6). It cannot be emphasized enough that “[a]ll quantitative models of language are wrong – but some are useful” (Grimmer & Stewart, p.3). For the research objective of the paper, which is devising an efficient way of building a classification of internationalization policies, adopting a quantitative view of language is useful because it allows me to considerably reduce the complexity of the texts. Moreover, I work under the assumption proposed by Hopkins and King that “[p]olicy makers (...) may be interested in finding the needle in the haystack (...), but social scientists are more commonly interested in characterizing the haystack.” (2010, p.230). Also, typologies are necessarily more useful in achieving the latter.
Fourthly, and finally for the time being, reducing data “serves the analysts’ need for efficient representations, especially of large volumes of data” (Krippendorff, 2004, p.84). Needless to say, this component of the research design is crucial in achieving the goal of any thesis or study. Thus, the method to reduce the data has to be chosen with the research objective in mind. In a seminal article on computer assisted content analysis methods for political texts, Grimmer and Stewart offer “an overview of text as data methods” (2015, p.2) which in the case of the current research project considerably helps to make the decision of what method to use for reducing the complexity of data (see Figure 1). For the moment it suffices to say that, in order to reach the objective of constructing a classification, my research project will be using Latent Dirichlet Allocation (LDA) to reduce the complexity of the data. The steps to making this decision and how they apply in my research practically will be discussed in the next subsection.

Figure 1: Overview of computer assisted content analysis for political texts and steps for data collection and analysis (Source: adapted from Grimmer & Stewart, 2015, p.2)

Steps in Data Collection and Analysis

---

10 The boxes colored in grey in Figure 1 represent the steps that apply to the research design of my thesis. They also help to follow the line of thought of the next subsection.
Once the theoretical decisions about the components of the research design have been made, two steps that are invariably part of any computer assisted content analysis – of any content analysis for that matter – follow: acquiring documents and preprocessing them. They are inextricably linked with the research objective of any study\textsuperscript{11}. As previously mentioned, the research objective of the thesis is to provide a classification of higher education internationalization policy. The categories of this classification remain unknown before the data analysis is conducted. As I have argued in the previous section on typologies, constructing the categories of the classification directly from empirical data avoids the problem of reification. Moreover, in cases where classificatory exercises were not previously conducted, as is partly true for internationalization, it is difficult to derive the categories in advance “due in part to the massive number of potential organizations of even a small number of texts” (Grimmer & Stewart, 2013, p.15). In turn, a way to arrive at such categories is to apply fully automated clustering techniques. These are a type of unsupervised learning methods that return a single cluster arrangement of the texts given as input and provide a way of extracting categories/dimensions of interest from them. Inverting for a second the chain of thought proposed by the \textbf{Figure 1}, I want to return to a previous proposition I made. I suggested that national policies for the internationalization of higher education express and operationalize a state’s understanding of the process. As such, I consider the position of the individual policies for internationalization in different clusters as being representative of the position of the state with regards to internationalization. As a result of this assumption the classification has to be carried out at the document level. In order to avoid the simplistic supposition that each document can be characterized by association with a single topic, the analysis proposes a to conduct the task of reducing data using Latent Dirichlet Allocation which is a method of clustering texts in mixed membership models.

To summarize, I distinguish between three essential steps in the process of reducing data: (1) acquiring documents, (2) preprocessing the documents, and (3) coding the documents using the LDA algorithm. These steps and how they apply to my research practically will be discussed in more detail next.

Firstly, in the context of my research acquiring documents involves systematically collecting national policies for the internationalization of higher education. Desk research soon revealed that

\textsuperscript{11} Even though in the current state \textbf{Figure 1} does not reflect this relationship.
there is no repository for national higher education internationalization plans or strategies and therefore, I had to build my own database. The question then became: how to carry out this task in a systematic manner? The solution proposed was to use the World Higher Education Database (WHED) built by the International Association of Universities which gathers systematic information about higher education institutions, systems and credentials worldwide. However, because their website is tedious to use, a web scraping\textsuperscript{12} exercise with Python\textsuperscript{13} was conducted. This endeavor made it possible to acquire an offline library of documents\textsuperscript{14} with systematic and reliable information about national higher education systems\textsuperscript{15} and national bodies responsible for international cooperation that proved invaluable for the next data collection step. I then proceeded to gather data for each country\textsuperscript{16} recording: (1) when the data was retrieved; (2) the websites of bodies responsible for governing higher education and international cooperation; (3) language(s) of instruction in higher education institutions; (4) remarks (on the country, higher education system, and policies) that I found pertinent to answering whether a higher education internationalization policy exists in the country; (5) existence of a national internationalization policy plan (split into 4 preliminary categories: yes, a section in general higher education strategy, a few sentences in national higher education strategy, no); (6) policy documents with higher education strategies (general or on internationalization), and reports and articles on the international dimension in the country’s higher education system. Giving a detailed account of the data gathered is beyond the scope of the current paper\textsuperscript{17}.

\textsuperscript{12} Also known as web harvesting or web data collection,
\textsuperscript{13} Python is an open source high level programming language like C++ or Java but with a clearer and simpler syntax structure.
\textsuperscript{14} When the data on credentials and systems was collected, the WHED had been last updated on the 6\textsuperscript{th} of June 2015.
\textsuperscript{15} The library has 206 entries: 189 countries, 13 provinces of Canada which are listed separately (Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Northwestern Territories, Nova Scotia, Nunavut, Ontario, Prince Edward Island, Quebec, Saskatchewan, Yukon), 3 territories of China (Hong Kong, Macao, Taiwan), and Belgium separated in Flemish and French Community.
\textsuperscript{16} This points to a question that is anything but trivial: How many countries are there in the world? Triangulating information did not work in this case as three fairly reliable sources (in terms of institutional development) came up with three different numbers: 189 in the WHED, 193 in the UN, and 209 in FIFA. I decided for the beginning to go with the number proposed by the WHED (as it was a data repository build for higher education so it fit with the research object of my thesis) and then also gather data on the UN members which were not in the WHED. As it turned out these were small islands that generally did not have fully fledged higher education systems and thus, no higher education policy. In the end the database I constructed has 198 entries.
\textsuperscript{17} The data recorded for points (1)-(5) for the 198 countries yielded a document of 46 pages (approximately 15,000 words).
Secondly, once the documents are acquired they have to be preprocessed for the analysis. This is a tedious, but very important step because of the “garbage in, garbage out” rule. To put it another way, if the quality of the data put into the analysis is not good, the results will not be either. This is hardly a specific issue of content analysis or computer assisted content analysis, all data collection methods (i.e. interviews, surveys, participant observation) have the same problem. The literature on computer assisted content analysis recommends a set of steps for preprocessing (Grimmer & Stewart, 2013) the documents: (1) transform .pdf documents into .txt documents; (2) clean the text (by removing: title page, content page, introductory arguments, executive summary, annexes/appendices, reference list/bibliography/ headers and footers); (3) select the text portion that deals directly with internationalization (as some countries have internationalization policy as part of a broader higher education policy); (4) apply UNIX encoding (there are special characters in texts that are otherwise not recognized by the computer); (5) tokenize the words (the process of breaking a stream of text into units, like words for example); (6) remove stop words (take out form the text the most common words in a language that provide not content, like “and”, “to”, “for”, “the”, “in”); (7) stem the words (the process of reducing the words to their root form, for example words like “international”, “internationally”, “internationalize”, “internationalization” become “internat*” ). Applying these procedures consistently to the documents that are to be analyzed is crucial if the results are to be reliable. Luckily, most of these tasks can be carried out by the computer program so full reliability can be ensured. So as to pilot test these procedures, a small pilot study with a convenience sample was conducted. The method and the preliminary results of this pilot test will be discussed next.

Thirdly, the method that fits the research objective of the study is applied so as to reduce the complexity of the data. As previously mentioned, this thesis will make use of Latent Dirichlet Allocation which is a “generative probabilistic model for collections of discrete data such as text corpora” (Blei, Ng & Jordan, 2003, p. 993). The goal of the algorithm is to:

“find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical

---

18 The pilot test with the proposed methodology was conducted on the higher national internationalization policies of Australia, Estonia, Finland, Germany, Japan, Malaysia, The Netherlands, New Zealand, Spain, and the UK.
relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments” (ibid.)

To put it simple terms, it is a way of automatically discovering topics from documents and thus, the unknown categories become known. This is achieved through three steps: (1) the researcher defines the number of topics/categories to be extracted from the text (through trial and error\(^\text{19}\)); (2) every word from each document is assigned (semi)randomly to a topic; and (3) the LDA algorithm updates this (semi)random assignment through an iterative process based on probabilities. As topic probabilities offer an explicit representation of the document, at the document level each document can be associated with multiple topics (what was previously referred to as mixed membership model). Because the statistical relationships are kept, documents (i.e. public policies for internationalization) can be easily clustered and compared with each other\(^\text{20}\). And, thus, countries and their higher education internationalization strategies can be compared with each other.

Three trials were conducted within the pilot study using LDA after preprocessing the documents (they are depicted in Figure 2). For the first trial the computer was given the task to find 3 topics from the texts and three words that would characterize these topics (see results in Figure 2). Remember, I did not tell the computer what the texts are about, but it seems to have figured out that they are about internationalization (a word that figures in all three topics), about education, research, universities and students. The words in each category seemed to be overlapping, probably because of the reduced number of texts given as input, so I decided to conduct a second trial with 2 topics and 4 characteristic words. As you can see from Figure 2 the topics have similar words, but in different proportions\(^\text{21}\). After seeing that the algorithm can correctly assess that the texts are about internationalization and education, I decided to optimize the program by feeding this piece of information into the algorithm\(^\text{22}\). Finally, I ran the program for a third time for 2 topics and 2 characteristic words for each. As can be seen from Figure 2, topic 1 is about students and universities and topic 2 is about students and research

\(^{19}\) This has to be informed by the type of data introduced, academic literature and, ultimately, by the judgment of the researcher. This issue will be discussed more when the results of the pilot study are presented.
\(^{20}\) My intuition is that this will prove valuable for assessing the validity of the results. Issues of validity are of paramount importance in computer assisted content analysis. However, they are beyond the scope of this paper which merely attempts to show the viability of the research design.
\(^{21}\) I did not put the proportions here, but I will explain why later on.
\(^{22}\) In essence, this meant to remove all the words with the stems “internat*” and “educ*”.
RESULTS:

TRIAL 1 (3 topics & 3 words):
TOPIC 1: educ* + intern* + student*
TOPIC 2: student* + educ* + intern*
TOPIC 3: intern* + research* + univers*

TRIAL II (2 topics & 4 words):
TOPIC 1: intern* + student + educ* + univers*
TOPIC 2: educ* + student + intern* + research*

TRIAL III (2 topics & 2 words):
TOPIC 1: student + univers*
TOPIC 2: student + research

Result: [u'0.030*student + 0.024*univers', u'0.019*student + 0.015*research']

The fact that students appear in both categories is not surprising either methodologically, nor theoretically. From the previous trials presented we can see that the word was picked up as relevant for multiple topics. From a theoretical perspective it is not surprising because international student mobility is considered to be one of the most important aspects of internationalization and so it can be expected that it would figure prominently in the higher education internationalization policies of the countries. The results also show that some policy documents on internationalization focus more on institutional internationalization, while others focus more on the internationalization of research. Again, internationalization literature supports this claim.

The proportions in which the words appear in each topic are also given for the third trial in the last line of Figure 2. I did not put the proportions for the other trials because as can be seen they are very small. This takes me back to the “garbage in, garbage out” rule. In the next steps I have to put more documents into the analysis and put more work into preprocessing the texts. For example, I adopted a very conservative list of stop words to be removed which I should probably extend or adapt to my purposes. I should also remove very rare words and words that I know appear often but are uninformative, for example “post-secondary”. All of these issues point out that there is more work to be done before the final data analysis exercise is conducted.
Conclusion

The internationalization of higher education remains a messy field, as only timid attempts were made to systematize the process (Kehm, 2003). Cleaning the vocabulary associated with the process of internationalization in higher education seems to be an insufficient exercise, since the concept is used in such a variable manner that it is questionable whether it can still serve as a valuable foundation for theory building. To put it another way, measuring internationalization and its impacts is difficult unless we know what we are actually measuring. The paper argues that the construction of a typology that combines the empirical and conceptual levels of analysis can represent a viable way of dealing with this problem as it would enhance the transparency of both scholarly research and policy practice. In turn, this would aid systematic knowledge accumulation that could better inform the design of policies aimed at forwarding internationalization in higher education. Moreover, the paper tried to tone down some of the most common point of criticism with regards to such heuristic tools by showing how these problems can be resolved. It also tried to describe the components of a computer assisted content analysis research design and show how the method of Latent Dirichlet Allocation might work in practice by showing some tentative results from a pilot study.

Applying this method brings about a number of contributions related to research design and methodology to the field of public policy. Firstly, it puts forward a novel methodology for analyzing not only higher education internationalization strategies, but also other kind of public policy documents from various fields. Considering the amount of procedural documents that are being produced by bureaucracy at every level, this type of methodology could prove invaluable for assessing issues like cooperation between departments. Secondly, it allows one researcher with limited resources (i.e. time, money, capacity) to investigate multiple cases. Thirdly, the methodology is perfectly replicable by other authors and with new policy fields. Fourthly, it provides perfect reliability of results as the same procedures/algorithms are applied consistently and systematically to all documents by the computer. Finally, the research will produce a database/repository for higher education policies for internationalization that could then be used by other researchers interested in the topic to explore different kinds of questions.
Bibliography


Berg, B.L., 2001, Qualitative Research Methods for the Social Sciences, 4th edn, Allyn and Bacon, Boston.


