Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention

Amelie Kutter and Cathleen Kantner

International Relations Online Working Paper, 2012/01

This paper is part of a series of method papers on the use of advanced text-analytical methods in the political and social sciences. The papers were presented at the two interrelated workshops "How to Analyse Millions of Texts and Still Be Home for Tea? Innovations in Textual Analysis in the Social Sciences" and "How to Make Use of Computer- and Corpus-Linguistic Methods for Large-n Text Analysis in the Social Sciences" conducted at Freie Universität Berlin in May 2010. The workshops were organised by Amelie Kutter and Cathleen Kantner of the FP6 Integrated Project "RECON – Reconstituting Democracy in Europe" and the Jean Monnet Centre of Excellence "The EU and its Citizens" at Freie Universität Berlin.

Amelie Kutter and Cathleen Kantner (2012): Corpus-Based Content Analysis: A Method for Investigating News Coverage on War and Intervention, International Relations Online Working Paper, 2012/01, February 2012, Stuttgart: Stuttgart University.

URL: http://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/

ISSN 2192-7278

The International Relations Online Working Paper series is intended to enhance scholarly as well as public debate on international politics and European issues. It invites papers from authors contributing to current debates in International Relations and European Studies in theoretically innovative ways or by the use of inventive qualitative or quantitative research methods. Interdisciplinary contributions on important international or EU-related issues are also very welcome. The International Relations Online Working Paper Series is published by the Department of International Relations and European Integration at the University of Stuttgart.

Papers are available in electronic format only and can be downloaded at http://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/

Issued by
Prof. Dr. Cathleen Kantner
Institute for Social Science
Department of International Relations and European Integration
University of Stuttgart
Breitscheidstr. 2
70174 Stuttgart, Germany
Tel: +49 (0) 711 / 6858 – 3425
Fax: +49 (0) 711 / 6858 – 3432

Contact: ir-working-papers@sowi.uni-stuttgart.de

The Authors

Dr. des. Amelie Kutter is researcher at the Department of Sociology at Lancaster University. She holds a PhD (European University Viadrina) and a MA (Free University Berlin) in Political Science and Cultural Studies. She has worked and lectured on issues of European integration and transnational communication, discourse and media studies, and methods of text analysis at Technical University Dresden, European University Viadrina Frankfurt (Odra), and Free University Berlin where she was member of the media analysis project presented in this paper.

Email: a.kutter@lancaster.ac.uk

Prof. Dr. Cathleen Kantner is professor of International Relations and European Integration at the Institute of Social Science at University of Stuttgart. She holds a PhD from Humboldt-University Berlin. She focuses on international relations, transnational public communication, European Security and Defence Policy and the emergence of transnational identities. Former teaching and research positions include Humboldt-University, Georg-August-University Göttingen, Bundeswehr Institute for Social Sciences in Strausberg, and Free University Berlin where she co-directed the media analysis project presented in this paper.

Email: cathleen.kantner@sowi.uni-stuttgart.de

Content

Abstract

Access to large digital text samples in various languages and the development of general-purpose tools in natural language processing holds great promises for comparative content analysis in the social sciences. However, boosted size, multiplicity of text genres and languages, and automation sharpen some classical problems inherent in the content analytical method. They reinforce the question of how certain (pre-defined) conceptual categories can be assigned to passages of text without violating the meaning that is implied by the entire text item through its linguistic structure.

The paper proposes *corpus-based content analysis* as a method to achieve high internal validity *and* increase in effectiveness through semi-automation. We suggest adopting the concept of 'semantic fields' as a heuristic tool for inference-making. It offers criteria that help to relate semantic information retrieved with the help of corpus-linguistic methods and computational tools to specific conceptual categories of a social science research project. An illustration of how to apply the method is provided for the conceptual categories: 'military intervention'; 'EU-Europe'; and 'actorness'.

Keywords: content analysis, corpus, semantic field, military intervention, method-mix, UK, France, US, Germany

1. Introduction

The last decade brought considerable advances in the digitalisation of data and the refinement of software for automated text analysis. This development has spurred research projects on political communication which employ large-n text samples (Grabowsky, 2011, Kantner, 2009, Kantner et al., 2008, Koenig et al., 2006, Kutter, 2011, Liebert, 2007, Renfordt, 2011, Trenz, 2004). They seek to put conventional content analysis on a broader data foundation. Using automated information retrieval and annotation techniques, they try to increase external validity across time, cases, and different languages. However, the boosted size of the text sample, the inclusion of diverse texts types, genres, and languages, as well as automated procedures sharpen some classic problems of the content analytical method, notably internal validity and reliability. The larger the sample and the less manageable the in-depth analysis, the more urgent is the question of how certain (pre-defined) conceptual categories can be assigned to a passage of text without violating the meaning that is implied by its linguistic structure and the context in which it occurred.

This dilemma became obvious during the comparative analysis of a large multilingual sample of newspaper articles that was carried out at Free University Berlin within the framework of the project 'In search of a new role in world politics. The common European foreign, security and defence policies (CFSP/ESDP) in the light of identity-debates in the member states' mass media'.[1] The compiled sample contained roughly half a million news articles on war and intervention that were published in selected print media in six European countries and the USA between January 1990 and March 2006.[2] Among other things, the analysis should reveal whether, in all the countries under investigation, the issue of humanitarian and military intervention was raised with similar intensity and brought into connection with the political

[1] The project was directed by Prof Thomas Risse and Prof Cathleen Kantner at Free University Berlin. It was funded by the German Research Foundation (DFG; contract no. RI 798/8)' and by the Sixth Framework Programme of the European Commission as part of the Integrated Project 'RECON – Reconstituting Democracy in Europe' (contract no. CIT4-CT-2006-028698) in the years 2005-2010. We wish to thank FAZIT foundation for sponsoring access to the archive of the *Frankfurter Allgemeine Zeitung* and *Rzeczpospolita* for having provided us with a tailored set of articles from the internal archive of the newspaper. We also wish to thank Barty Begley for the language editing.

[2] The corpus includes articles published on war and military intervention during the years 1990-2006 from: the *New York Times*, *The Washington Post* (USA); *The Guardian, The Times* (UK); *The Irish Times* (Ireland); *NRC Handelsblad, De Volkskrant* (Netherlands); *Le Monde, Le Figaro, Les Echos* (France), *Süddeutsche Zeitung, Frankfurter Allgemeine Zeitung* (Germany); *Der Standard, Die Presse* (Austria). The Polish newspapers *Gazeta Wyborcza* and *Rzeczpospolita*, representing the seventh European country, were included in the corpus linguistic procedures, but not in subsequent counting, due to lacking meta-data.

space 'Europe' and with conceptions of the European Union (EU) as an international actor. If this was the case – this was the assumption derived from Habermasian conceptions of European public spheres –, one could reasonably expect that transnational debates on shared ethical standards of military intervention and the international role of the EU had emerged (Kantner, 2009, Kantner et al., 2008). In addition, such debate was likely to provide (critical) resonance for the EU's Common Foreign, Security and Defence Policies and their further institutionalisation (Kantner and Kutter, 2011).[3] But how could we establish whether military action was actually conceived and discussed as (ethically justified or justifiable) 'intervention', rather than a conventional war; whether the European political space (countries, actors, policies) was concerned, rather than the broader geographical space; and whether the EU was portrayed as being in charge and as having/lacking the capacity to act on these events?

With half a million newspaper articles to hand, the default method of text analysis in the social sciences – content analysis – was reaching its limits.[4] Content analysis is a standardised hermeneutic procedure of text interpretation in the course of which the individual analyst assigns abstract categories to propositional contents (issues, claims, problem definitions, frames) that occur in passages of the analysed texts. The categories do not correspond to the 'observable' (linguistic) characteristics of the texts, but to hypotheses derived from social theory about the social and political setting in which these texts occur (e.g. the change of problem perceptions and collective identities after the Cold War). The content analyst tries to make replicable and valid abductive inferences from texts to these conceptual categories (Krippendorff 2004: 18). The validity and reliability of assignment (coding) is assured by a classification system (codebook). The codebook details which hypothesis relates to which content dimension and which content variants typically and expectably embody this dimension in the texts to be investigated. Having classified texts accordingly, the analyst disposes of meta-data that can be used for theory-building (if following a Grounded Theory paradigm) or theory-testing (if following a more empiricist set of assumptions) by means of descriptive or analytical statistics.

The content analysis community has developed sophisticated measures for controlling codebook compliance and convergence of coding decisions (intercoder-reliability, cf. Holsti,

[3] Other research questions concerned the emergence of (similar) interpretative frames and normative-ethical judgements about 'intervention' e.g. with regard to communitarian or pragmatic conceptions of community, international law (Renfordt, 2010, 2011), or experience of national involvement in military and humanitarian interventions (Grabowsky, 2011).

[4] However, a qualitative content analysis was applied to a sub-set of 6,000 randomly sampled articles on intervention with the help of a large team of trained coders (for some of the results see Grabowsky, 2011, Kantner et al., 2008, Renfordt, 2010).

1969, Krippendorff, 2004). Yet, the validity and reliability of making inferences from concrete texts is somewhat neglected, i.e. the question how 'typical occurrences' are to be defined so that they can be replicably recognized by human coders and, importantly, computers as 'indicating' a particular (non-observable) category (Weber, 1990). As a rule, codebooks do not define which structural features of texts such as lexical or syntactic units should be considered as representing a particular content dimension, and under which conditions.

In 'qualitative' content analyses of smaller text samples, this problem is somewhat balanced: for inference-making, the individual reader implicitly employs her knowledge about the entire text item, her linguistic competence, professional background knowledge and training (Mayring, 2002). Still, implicitness means lack of transparency and replicability: we do not know how the individual coder has dealt with ambiguity of words and variability of terms while making recurrent coding decisions, how she made a text "appear to 'speak' in the analytical terms that the (…) analyst is familiar with" (Krippendorff, 2004: 34). In quantitative approaches, this problem is even more acute as information is usually exclusively inferred from single words (considered as indicative of the social problem in question), after brief consultation of 'Key Word in Context' displays, i.e. the display of the immediate co-text of a keyword (e.g. in the software CAQDAS or QDA Miner).

The risks of skewed indication and representation resulting from content analysts' indifference towards linguistic features of texts have long been discussed (Fühlau, 1982, Knapp, 2008, Roberts, 1989). What is missing to date is an "adequate theory of language to direct us in finding the alternative signs that express a particular concept." (Stone, 1966: 9f). All the more promising, therefore, appear tools of natural language processing (NPL) that, drawing on linguistic theories, take complex syntactic and semantic information into account to represent meanings of texts. This information is usually automatically retrieved from text collections that had been 'tagged' before with information on the grammatical and syntactical use of words or with human readers' interpretations. The availability of these technologies has nurtured the hope among social scientists that there is an easy fix to the content analysis of large-n text samples: press the button of some text-mining tool, and your content analysis will be done in two seconds!

However, the available tools still yield a (fragmentary) linguistic reading of the texts. As a rule, they produce information on the specific vocabulary and generic characteristics of texts instead of issues, proposals, and frames investigated in a social science text analysis. Most NPL tools classify texts according to semantic regions that are either too general (e.g. the differentiation into 'economy', 'geography', 'politics' in WordNet) or too narrow (e.g. taking proper names as indicator for 'actor') to correspond to the specific research question which a

social scientist might develop. Efforts at modelling NPL on content-analytical criteria tend to approximate only some aspects of the abstraction performed by a human reader and need additional manual input for improved accuracy (Grimmer and King, 2011, Wüest et al., 2011). Hence, the challenge remains to find a transparent and replicable method of capturing linguistic units (such as words and word-clusters) so that they are (a) representative of the actual text and meaning they are part of; and (b) indicative of the conceptual category selected by the social science researcher for qualitative-quantitative study.

This paper suggests a methodology for qualitative-quantitative inference-making that could be named *corpus-based content analysis* and is applicable in the field of semantics. We argue that the concept of the 'semantic field' helps to understand the ambivalent correspondence between lexical-syntactic units on the one hand, and abstract concepts on the other. Further, categories and tools from corpus linguistics and computational linguistics, which seize the syntagmatic constitution of semantic clusters, facilitate the identification of term variants and word senses that approximate abstract social science concepts. Once identified, the lexical approximations of a social science concept can be searched, and the results can be used for a descriptive or analytical-statistic assessment of research questions in the manner in which content analysis usually proceeds.

We will illustrate this technique with examples from the large multilingual corpus of news coverage on war and military intervention gathered within the framework of the project.[5] After cleaning, this corpus comprised 489,508 articles and 393,268,544 words. It was tokenised and annotated and then subjected to manual analysis in WordSmith and automated analysis in 'DISCO', an application developed by Peter Kolb for the representation of semantic spaces (Kolb, 2008). The paper benefited from collaboration with Manfred Stede and Peter Kolb[6], from student assistants' involvement with natural language processing[7], and from the (foreign) language skills of five human concordance analysers who participated in language-contrasting team-sessions on disambiguation.[8]

---

5 A *corpus* is a collection of texts assumed to be representative of a given language or sub-set of language (e.g. spoken vs. written texts, newspaper vs. academic texts) or of a thematic focus (e.g. humanitarian military intervention) that has been prepared for computer-aided corpus analysis.

6 Department of Linguistics, University Potsdam.

7 Andreas Hildebrandt (Department of Linguistics, University Potsdam); Mark Püttcher (Otto-Suhr Institute, Free University Berlin).

8 We wish to thank Dominika Biegoń, Anaïs Bordes, Jana-Katharina Grabowsky, Joshua Rogers, and Ewelina Sokołowska (all Otto-Suhr Institute of Political Science, Free University Berlin at the time of project-involvement) for their meticulous concordance analyses. Many thanks for feedback also to Christiane Hümmer, Majid Khosravinik, and Paul Sarazin.

The paper is structured as follows: Section 2 below seeks to familiarise the reader with the major assumptions of semantic field theory and introduces corpus-semantic tools that reveal semantic clustering. Section 3 sets out the procedures that we developed drawing on these insights. In section 4, we will discuss three examples of application: the concept of 'military and humanitarian intervention'; 'EU-Europe'; and 'actorness' (section 4). Finally, section 5 will draw some conclusions as regards 'rules of thumb' for corpus-based content analysis.

2. Semantic fields and corpus semantic methods

Semantic field theorists start from the structuralist assumption that the meaning of a single word or linguistic unit is determined by paradigmatic relations (substitution or opposition) and syntagmatic relations (hierarchy or sequence) with other words that belong to the lexicon of a natural language (Trier, 1931, Ullmann, 1951, Vassilyev, 1974). They claim that the vocabulary of a natural language is "internally structured by many clusters of words, which stand in different relations to each other, sometimes logical relations of sameness, difference and entailment, sometimes vaguer relations within a topic area or semantic field." (Stubbs, 2001: 35). The meaning of a single word is, accordingly, constituted by its (paradigmatic and syntagmatic) relations to other words that *cluster semantically*, i.e. that refer to the same partial *conceptual region* of the vocabulary of a natural language: a semantic field (Gliozzo and Strapparava, 2009: 15). A semantic field consists "of basic key-words, which command an army of others. The semantic area may be regarded as a network of hundreds of associations, each word of which is capable of being the centre of a web of associations radiating in all directions. A word like 'man' might have as many as fifty such associations – chap, fellow, guy, gentleman, etc." (Mackey 1965: 76). A semantic field can be understood as "a closely knit and articulated lexical sphere where the significance of each unit is determined by its neighbours" (Ullmann 1951: 157).

Yet, a semantic field is never identical with the linguistic units it is composed of – these may refer simultaneously to various different conceptual regions of a language (word sense ambiguity). Moreover, a semantic field is not fixed to particular terms, but may be constituted varyingly by different terms (variability) (Gliozzo and Strapparava, 2009: 16). Disambiguation of word sense and identification of variants, two preconditions of text comprehension, are consequently entirely dependent upon context. They are dependent upon the specific lexicon of a language that a speech community is acquainted with (as lexical field theory suggests), the usages of speech and writing in a particular pragmatic situation (as Wittgenstein's semantics would suggest), and the distribution of syntactic positions within a set of texts from

a particular domain of social activity (distributional semantics) (Gliozzo and Strapparava, 2009, Kilgariff, 2003). Hence, the reconstruction of a conceptual region of a natural language requires the careful examination of (historically specific) language use.

At the same time, semantic fields are schemes that have default values, typical realisations, and associations which can be abstracted from lexical realisations. They provide for 'norms' in social communication and 'mental models' in individual text comprehension across languages (Stubbs, 2001: 96). As abstract entities, semantic fields may converge across different natural languages even though the respective linguistic units do not correspond to each other – an insight applied, in particular, in translation studies and contrastive lexicography. Corpus semantic studies draw on semantic field theory. However, whereas the founding theoreticians of semantic fields used introspection (judgement based on the analyst's knowledge about linguistic structure) to identify conceptual regions of a natural language (mainly in their paradigmatic aspect), corpus approaches pertain to an evidence-based description of language (mainly in their syntagmatic aspect). This description is based upon primary and secondary (annotated) data which has been retrieved from large electronic corpora by means of computer-aided methods (Lenz, 2000). Frequent patterns of co-occurrence, syntactic sequence and hierarchy of words are seen as indicating semantic clustering. This assumption follows from "a contextual theory of meaning, where a text is seen as an integral part of its context and the formalization of contextual patterning of a given word or expression is assumed to be relevant to the identification of the meaning of that word or expression" (Tognini-Bonelli, 2001: 4). The assumption that lexical-contextual patterning indicates semantic clustering has informed both *manual* corpus-linguistic studies and *automated* procedures of natural language processing.

Corpus-linguistic studies focus on the computer-aided exploration of lexical, syntactical, or pragmatic triggers of semantic similarity and relatedness which occur in a specific text collection (Baker, 2006, Simon-Vandenbergen and Aijmer, 2002, Stubbs, 2001). Following Stubbs, features of semantic relatedness can be modelled "as clusters of lexis (node and collocates), grammar (colligation), semantics (preferences for words from particular lexical fields) and pragmatics (connotations or discourse prosodies)." These relations are defined as follows (Stubbs, 2001: 24; 65):

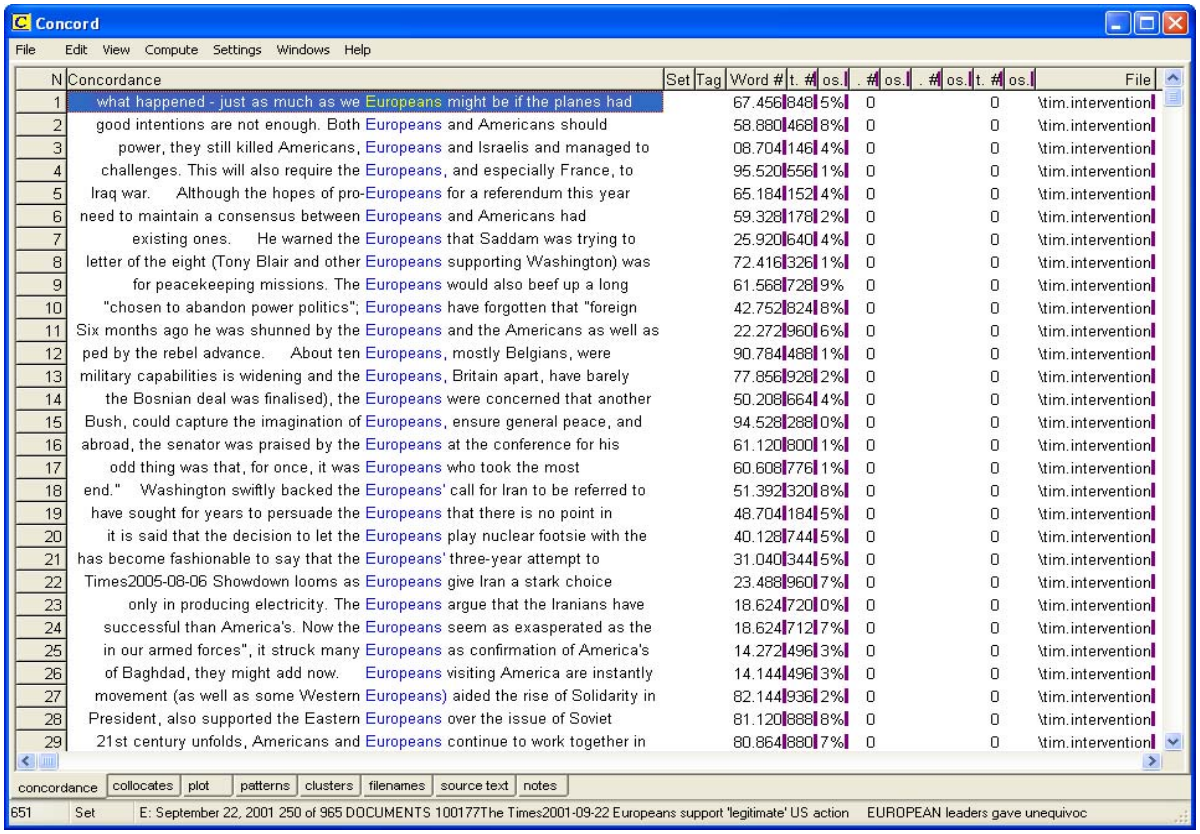
- Collocation is the "lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text."

- "Colligation is the relation between a pair of grammatical categories or, in a slightly wider sense, a pairing of lexis and grammar. For example, the word-form *cases* frequently co-occurs with the grammatical category of quantifier (some, many, more, both, several)."
- "Semantic preference is the relation, not between individual words, but between a lemma or word-form and a set of semantically related words, and often it is not difficult to find a semantic label for the set." An example is the word-form *large*, which often co-occurs with words for quantities and sizes.
- Discourse prosody is a feature extending over more than one unit in a linear string, which implies evaluation and expresses speaker's attitude. For instance, the lemma 'cause' co-occurs primarily with unpleasant events, while the lemma 'provide' occurs with words denoting things which are desirable or necessary.

Such syntagmatic semantic clustering can be explored with the help of displays of words like wordlists or concordances provided by corpus-linguistic software such as WordSmith. *Wordlists* sort all the words or lemmas occurring in a corpus according to frequency of occurrence. They reveal potential node words that are representative for the field or subject covered by the corpus. When comparing the wordlist of one specific corpus to wordlists of reference corpora (e.g. a general natural language corpus such as the BNC), one will get hold of those words that are comparatively overrepresented or underrepresented ('key words'). Such comparison reveals the distinct (the 'key') lexical characteristics of the corpus, its (genre-)specific vocabulary. However, word-form types displayed in wordlists are not "considered to be the best format for studying the lexicon in a corpus for the obvious reason that polysemy and word-class ambiguity cannot be distinguished. For this reason, the procedure known as concordancing, providing the context in which each word token appears, has been the major tool used for accessing corpora." (Kennedy, 1998: 247).

A *concordance* is the formatted version or display of all the occurrences of a particular type (i.e. the search word or word cluster) in the corpus and of the words regularly co-occurring with the search item as well as whole clusters of words of which the search-word is a part on a regular basis. The most usual format is *Key Word in Context* (KWIC) which displays the concordances of a particular key word ('node') in such a way that the same word sequences to the right or the left of the key word are displayed together ('node-collocate-span') (Kennedy, 1998: 251-256; see Figure 1).

Figure 1: KWIC View of 'Europeans' in British News Articles on 'Intervention'



Note: Figure 1 is a screenshot of the most frequent occurrences of 'Europeans' and their immediate co-text in the British newspaper articles on humanitarian and military intervention as displayed in WordSmith.

The words that co-occur with one particular search word can also be displayed with regard to particular features of regularity, e.g.: according to their overall frequency of co-occurrence ('collocates' view in WordSmith); according to the syntactical position in which they co-occur most frequently with the search word ('patterns' display in WordSmith); or according to 'adjacency', displaying those words that are most frequently adjacent to the search word ('n-grams'; 'clusters' view in WordSmith). These displays provide information on the company words keep in a specific corpus and, hence, information that helps to judge the semantic relations of a search word in that corpus.

NPL-based semantic studies centre on the identification and display of 'semantic domains' through the automated processing of large text corpora. Semantic domains are semantic fields that are characterized by lexically coherent words whose main property is to co-occur in texts (Gliozzo, 2006: 5). They can be classified either according to semantic similarity of lexical concepts or according to distributional similarity of words occurring in a corpus. The former method draws on the lexical coherence assumption (*lexical concepts* occurring in a text tend to belong to the same semantic domain) and relies upon comparison with given,

hand-crafted lexicons or classification system on domains of human activity or discussion (e.g. politics, economy, geography in WordNet) (Gliozzo and Strapparava, 2009: 4f, Magnini et al., 2002). The latter method is based on the distributional similarity or relatedness hypothesis (*words* occurring in the similar contexts tend to have a similar meaning when they have many common co-occurring words in the same syntactic relations or in the same distributional context). It models semantic domains by comparing the syntactic position and distribution of words within a given set of texts (Kolb, 2008). Information on the lexical coherence or distributional similarity of words can then be used for word sense disambiguation and concept-extraction, cross-lingual lexicons and ontology-generation.[9] The following section will show how one can make use of the insights from semantic field theory and corpus semantics for a corpus-based content analysis of large multilingual text samples.

3. Procedures of corpus-based content analysis

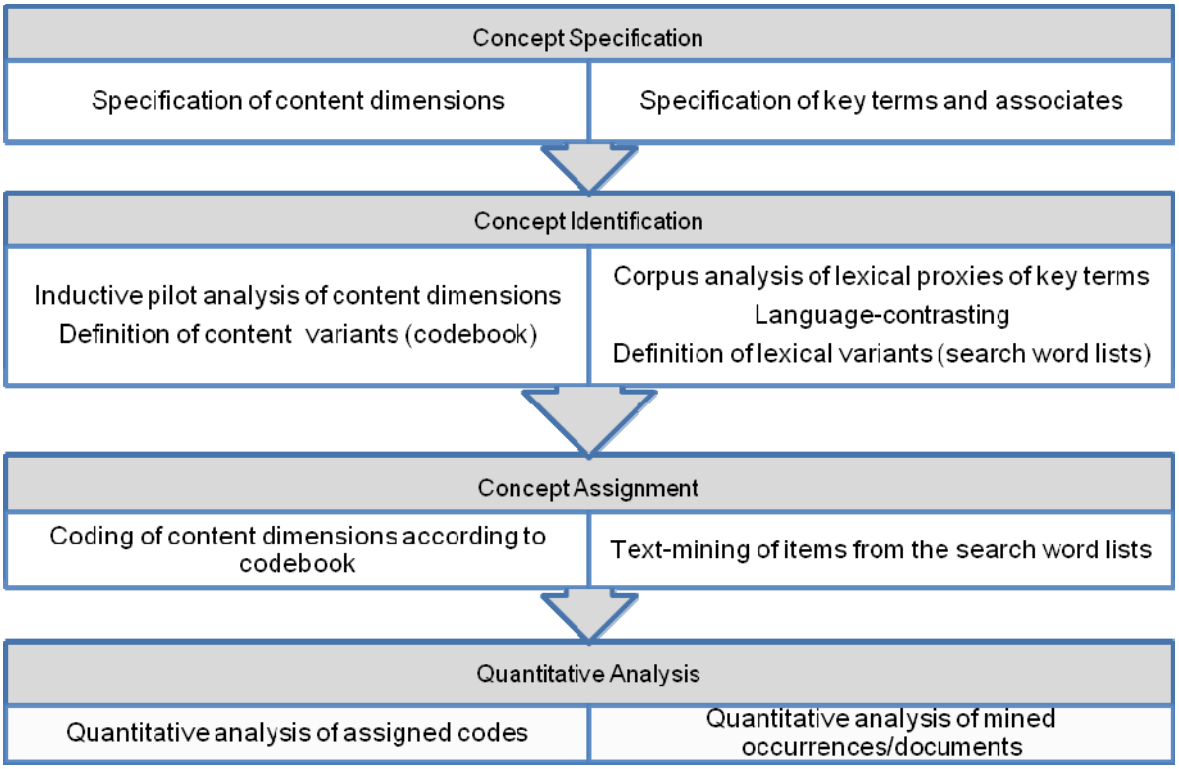
The major insight of semantic field theory for corpus-based content analysis lies in the following assumption: the meaning of a word is given by its belonging to a conceptual region of a natural language which, in turn, is constituted through semantically related words. To some extent, this supposition can be adopted for social science concepts. They, too, are constituted through key terms and semantically related terms. Even though they might not coincide with lexical fields proper as occurring in a natural language, they may be approximated by a set of lexical items or correspond to a set of lexical fields. Signal words like ‘troops’, for instance, might associate semantic clusters which meaningfully relate to the political science concept of ‘intervention’. Having accepted this admission, one can derive more specific suggestions from semantic field theory for concrete inference-making. Firstly, inference-making that mediates between text and conceptual category can be systematised according to types of relations between words: according to paradigmatic relations, which indicate degrees of similarity, dissimilarity, and entailment (synonyms, antonyms, hyponyms, meronyms); and according to syntagmatic relations of sequence and syntactic positioning. Secondly, such systematisation has to be aware of word sense ambiguity (the many meanings a word can have) and term variants (the lexical realisations a concept can have) and their context-dependent disambiguation, selection or clustering. Thirdly, once patterns of semantic clustering are identified in the respective languages, they can be abstracted from the lexical realisations and contrasted across the different languages. In short, the concept of

[9] In information science, ‘ontologies’ are formal representations of shared conceptualisations of knowledge.

the semantic field provides a link missing so far in content analysis: a heuristic tool which relates concepts to textual structure.

The methods and tools of corpus semantic studies described in the previous section support such inference-making. They reveal the syntagmatic relations a specific search term has in a specific corpus. NPL tools generate overviews over term variants, word senses, and associates of a search term in a specific corpus, and hence give an idea of the lexicon this search term associates. Corpus linguistic procedures (wordlists, concordances) allow for additional inductive reconstruction of lexical items or lexical fields that correspond to a social science category. Along with the semantic field theory, these tools and methods help to 'turn' content analysis into a corpus-based procedure that is adequate for the quantitative analysis of large text samples. Figure 2 shows the procedure we developed in the course of the project as compared to conventional content analysis. The procedure retains the steps that are usually taken in a content analysis, but orients them on semantic field theory and corpus analysis.

Figure 2: The Content Analytical Procedure (Conventional vs. Corpus-Based)



In the step 'concept specification', the preferred notion of the investigated social science concept (e.g. 'intervention', 'actorness') is established, drawing on relevant theoretical literature. While conventional content analysis would make out the content dimensions of the concept, the corpus-based approach would focus on those terms which, according to social

science literature, are constitutive for and indicative of the concept that is being investigated. For 'concept identification', conventional content analysis usually starts with an inductive pilot analysis in order to figure out which content variants typically represent the selected content dimensions. This may result in a reformulation of the investigated concept in the light of the pilot study. Based on this initial exploration, one would construct a codebook and train the 'coders' in corresponding manual (computer-aided) annotation.

The corpus-based version of content analysis proceeded differently for concept identification. We searched for lexical items that came closest to the investigated social science concept. Using corpus and computational linguistic tools, in particular WordSmith and DISCO, we tried to establish term variants and word senses for each natural language involved in the project. The search consisted of four subsequent moves:

- searching existing (corpus-based) thesauruses for term variants and word senses that overlap with the social science concept specified before, for initial orientation as to which lexical units might be relevant;
- searching the corpus under investigation for (a) the salience and keyness of terms indicative of the concept, using wordlists in WordSmith; (b) the semantic clusters these terms had in the corpus, using concordance analysis in WordSmith or DISCO;[10]
- contrasting these findings across languages and searching for equivalents of expressions that proved salient in one language, but not in another;
- inclusion of all those term variants and node-collocate-spans into a *list of search words* that, in all languages, were unambiguously related to the investigated concept; exclusion of term variants and word clusters that were related to not only the investigated concept or that do not relate to it at all.

'Concept assignment', the third step, which takes most effort and time in the conventional version of content analysis, is automated in the corpus-based version. All texts under investigation are 'mined' with respect to whether or not they contain one or several of the lexical items that had been included in the search word list before (but not the excluded lexical items). This can be done by connecting character strings, i.e. the line of adjacent characters that make up a word or a node-collocate-span from the search word list, through Boolean operators; and through searching ('mining') occurrences of these character strings in the texts under investigation. Software like SPSS Clementine, the (semi-)automated coding function in Atlas.ti, or the match list function in WordSmith provide ready-made

[10] This step is very similar to the method proposed in relation to a corpus-based discourse study of press coverage on refugees (Baker et al., 2008, Baker and McEnery, 2005).

solutions for such simple text mining. However, depending on the specific research project, it is better to replace them with a tailored programme.[11] One way or the other, one arrives at the number of occurrences of one or several lexical realisations of the investigated concept-category. As primitive as this might be compared to more the complex text mining in advanced NPL tools, this simple procedure ensures that occurrences of only those character strings are matched and counted that had been identified before as unambiguously representing the investigated concept-category. In the fourth step, the information retrieved and quantified in that way can be used for the quantitative analysis of frequencies, occurrences or correlations in exactly the same manner as codes are investigated in conventional content analysis. The following section will demonstrate how we applied the corpus-based content analytical procedure to major concept-categories of the project: 'intervention', 'EU-Europe', and 'actorness'.

4. A corpus-based content analysis of news on war and intervention

The objective of our content analysis of media coverage on war and intervention was to clarify the relation between the transnational occurrence of media debates on international conflict resolution, reference to 'EU-Europe' as affected or involved political entity, and the development of (similar) interpretative frames and normative-ethical judgements about 'intervention' relating e.g. to communitarian or pragmatic conceptions of community (Kantner, 2009), international law (Renfordt, 2010, 2011), or historical references of national involvement (Grabowsky, 2011). Moreover, we were interested in revealing whether the European Union, when referred to in relation to violent conflicts, was constructed as an international actor that ought to act upon international crises (Kantner et al., 2008) and, if so, whether this related meaningfully to the development of the EU's foreign, security, and defence policies (Kantner and Kutter, 2011). To this end, various specific questions had to be settled, e.g.: when military action was actually conceived and contested as (morally justified or justifiable) 'intervention'; whether the European political space (countries, actors, policies) was seen as concerned by these events; and whether international 'actorness' was attributed to the EU.

[11] The project started out using SPSS Clementine, but switched to tailored software programmed by Andreas Hildebrand since the text mining was entirely intransparent in SPSS Clementine and could not be manipulated in the required way.

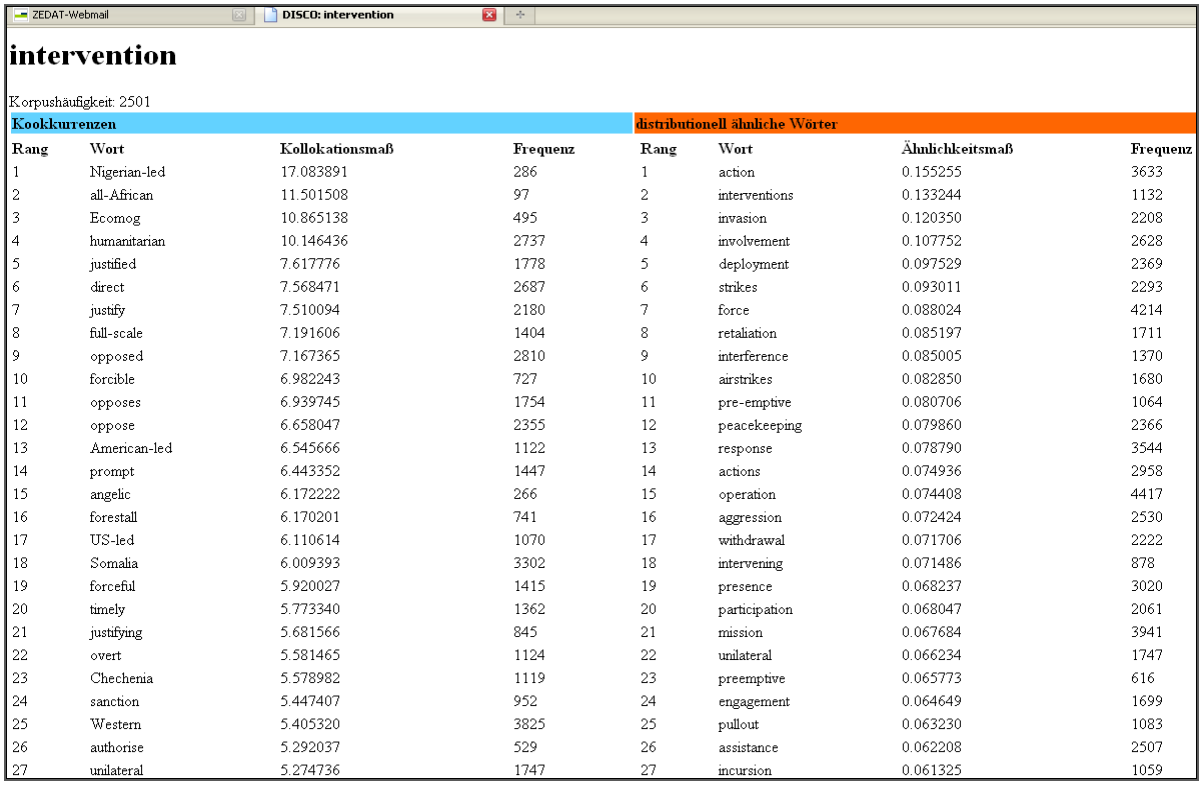
4.1. Identifying term variants and news salience of 'intervention'

The research question related to the category 'intervention' was thus: what news articles and how many of them issue the problem of 'military and humanitarian intervention' at what time? The corpus had been sampled from web archives running a search for articles that contained the words 'war' or 'intervention' in combination with names of regions and countries where conflicts had taken place during the period under investigation (1990-2006). The task was to isolate those news articles, without reading them in full, which in some way related to the issue 'intervention' as conceived in International Relations and International Law. From a foreign policy perspective, 'intervention' denotes a state's or an alliance's military action in the territory of a third party, where a violent conflict is about to escalate, with the purpose of either supporting one of the fighting parties or decreasing the fighting. It may also denote the deployment of troops in order to settle a humanitarian catastrophe and to provide humanitarian aid.[12] As it affects the sovereignty of the intervened states or regions and involves authorisation of multilateral action, 'intervention' is also a subject of international law. International law stipulates conditions of intervention, such as the consent of the government in a conflict region or the decision of the Security Council, and specifies procedures and terms of action to be followed. Hence, differently from ordinary 'war', 'intervention' pertains to more sophisticated moral-legal justification (Kantner, 2009). Accordingly, humanitarian and military intervention is likely to be associated with specific expressions of the lexical field of military action (strikes, attacks) and military operation (troops, forces, deployment), with the terminology of UN conflict management (peacekeeping, stabilisation mission, UN resolutions etc.), and particular missions (KFOR, UNIMOC etc.).

However, media agents are unlikely to use the denotations of professional academic or legal discourse. Moreover, when compiling the news articles from the web, we had realised that military action was not necessarily framed as 'intervention' even if it was an intervention according to the definitions of international relations and international law- Instead, wording used to describe ordinary 'wars' (unilateral occupations, armed conflicts) was frequently employed (Kantner et al., 2011, Kantner and Renfordt, 2007). Thus, the difficulty was to identify media-specific mentions of 'intervention' that were not designated as such.

[12] Source: http://de.wikipedia.org/wiki/Intervention (21.12.2006).

Figure 3: Display of Semantic Relations of 'Intervention' in DISCO



ZEDAT-Webmail | DISCO: intervention

intervention

Korpushäufigkeit: 2501

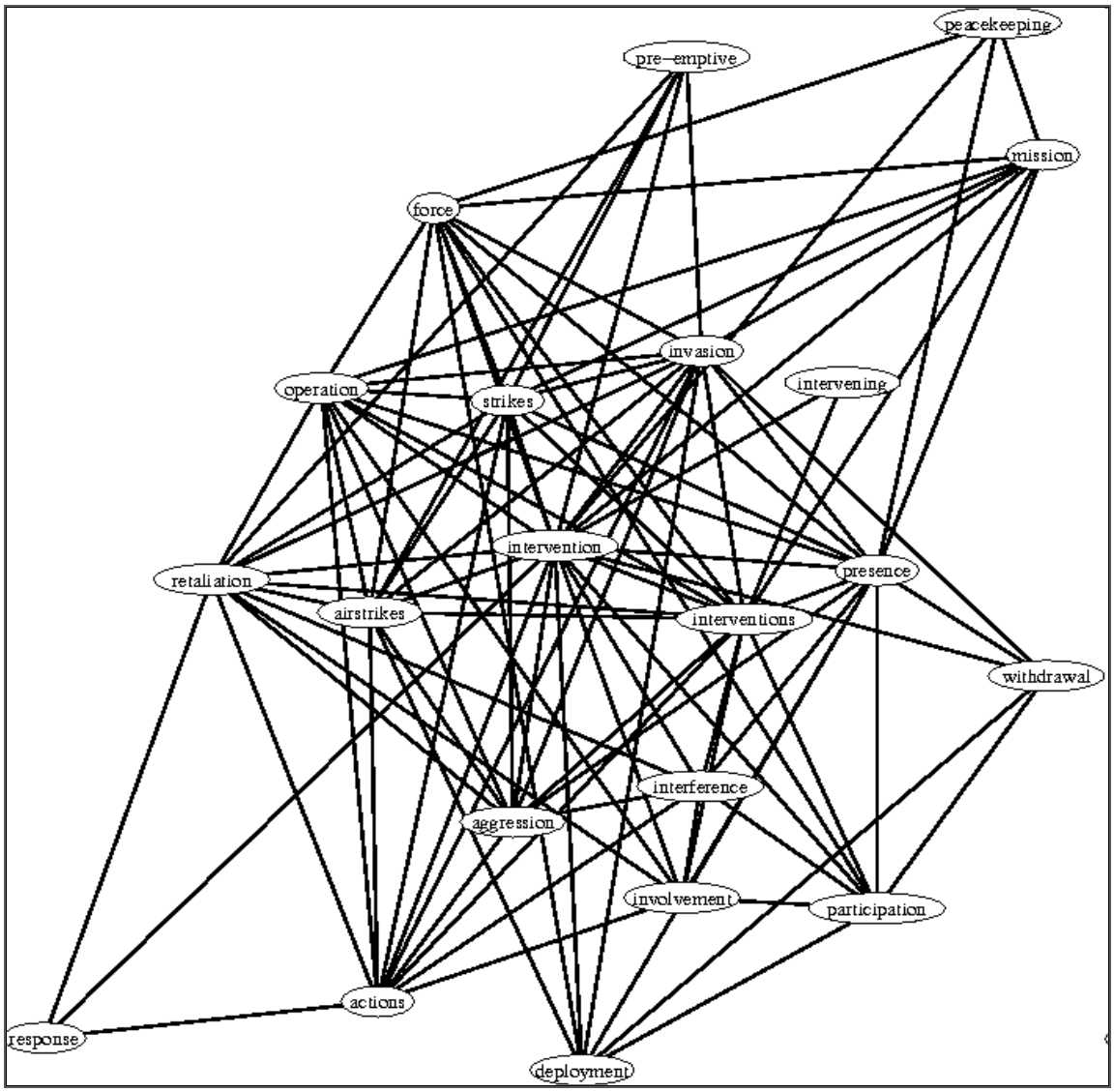
Kookkurrenzen				distributionell ähnliche Wörter			
Rang	Wort	Kollokationsmaß	Frequenz	Rang	Wort	Ähnlichkeitsmaß	Frequenz
1	Nigerian-led	17.083891	286	1	action	0.155255	3633
2	all-African	11.501508	97	2	interventions	0.133244	1132
3	Ecomog	10.865138	495	3	invasion	0.120350	2208
4	humanitarian	10.146436	2737	4	involvement	0.107752	2628
5	justified	7.617776	1778	5	deployment	0.097529	2369
6	direct	7.568471	2687	6	strikes	0.093011	2293
7	justify	7.510094	2180	7	force	0.088024	4214
8	full-scale	7.191606	1404	8	retaliation	0.085197	1711
9	opposed	7.167365	2810	9	interference	0.085005	1370
10	forcible	6.982243	727	10	airstrikes	0.082850	1680
11	opposes	6.939745	1754	11	pre-emptive	0.080706	1064
12	oppose	6.658047	2355	12	peacekeeping	0.079860	2366
13	American-led	6.545666	1122	13	response	0.078790	3544
14	prompt	6.443352	1447	14	actions	0.074936	2958
15	angelic	6.172222	266	15	operation	0.074408	4417
16	forestall	6.170201	741	16	aggression	0.072424	2530
17	US-led	6.110614	1070	17	withdrawal	0.071706	2222
18	Somalia	6.009393	3302	18	intervening	0.071486	878
19	forceful	5.920027	1415	19	presence	0.068237	3020
20	timely	5.773340	1362	20	participation	0.068047	2061
21	justifying	5.681566	845	21	mission	0.067684	3941
22	overt	5.581465	1124	22	unilateral	0.066234	1747
23	Chechenia	5.578982	1119	23	preemptive	0.065773	616
24	sanction	5.447407	952	24	engagement	0.064649	1699
25	Western	5.405320	3825	25	pullout	0.063230	1083
26	authorise	5.292037	529	26	assistance	0.062208	2507
27	unilateral	5.274736	1747	27	incursion	0.061325	1059

Note: the figure displays the word company which the search word 'intervention' keeps in the English language newspaper articles of our corpus. It ranks collocates according to frequency of co-occurrence (left-hand blue-coloured column) and lists those words that, compared to 'intervention', occur in similar syntactic positions according to the degree of distributional similarity (right-hand red-coloured column).

A first approach to identifying term variants indicative of the category 'military and humanitarian intervention' was to process an analysis in DISCO. The tool counts both the frequency of lexical collocates of a node (a search word) and of those words that, compared to the node, are similarly distributed in the corpus.[13] Applied to a non-specific sample of texts available, for instance, at Wikipedia, this application will reveal different domain-specific versions of 'intervention' used in the field of economics, sports, politics, etc. When applied to our specific corpus, which we had cleaned beforehand of 'erroneously sampled' articles (e.g. from articles entailing uses of 'war' and 'intervention' in sports, cf. Kantner et al., 2011), the tool produced a detailed set of terms relating to the issue of 'intervention' in terms of international relations and military action. DISCO displayed these terms either in the form of a list (see Figure 3) or in the form of a network graph (Figure 4).

13 A simple context window of size ±3 words was used for counting co-occurrences, which took the exact position within the window into account. "This can be seen as a crude approximation of syntactic dependency relations. Instead of dependency triples like <donut, OBJ-OF, eat> we get triples of the form <donut, -2, eat>. Consequently, the features that describe a word's distribution are not just words as in a pure bag-of-words approach, but ordered pairs of word and window position" (Kolb 2008: 6).

Figure 4: The Semantic Space of 'Intervention' according to DISCO



Note: the graph displays those words in the English language newspaper articles of our corpus that were in syntactic positions most similar to those of the search word 'intervention'. Due to their distributional similarity, they are regarded as partial synonyms and hyponyms of 'intervention'. The graph displays an arbitrary snap shot of computation and does not bear information on the relatedness of the terms other than that they co-occur with 'intervention' in the corpus.

The frequency analysis of collocations is seen here as representing the syntagmatic perspective on the semantic cluster of 'intervention': it reveals the immediate collocates of 'intervention', such as the attribute 'UN-sanctioned' or 'forceful', the verb 'oppose' and 'justify', or the proper names 'Chechnya' and 'Somalia' (see column 'Kookurrenzen' in Figure 3). The distribution analysis of syntagmatic positions of words, on the other hand, is seen as representing the cluster from a paradigmatic perspective, showing partial synonyms like 'invasion' and/or hyponymous semantic fields of military action and operation such as

‘(air) strikes’, ‘deployment’, ‘withdrawal’, ‘pullout’ etc. (see Figure 4 and the column ‘distributionell ähnliche Wörter’ in Figure 3).

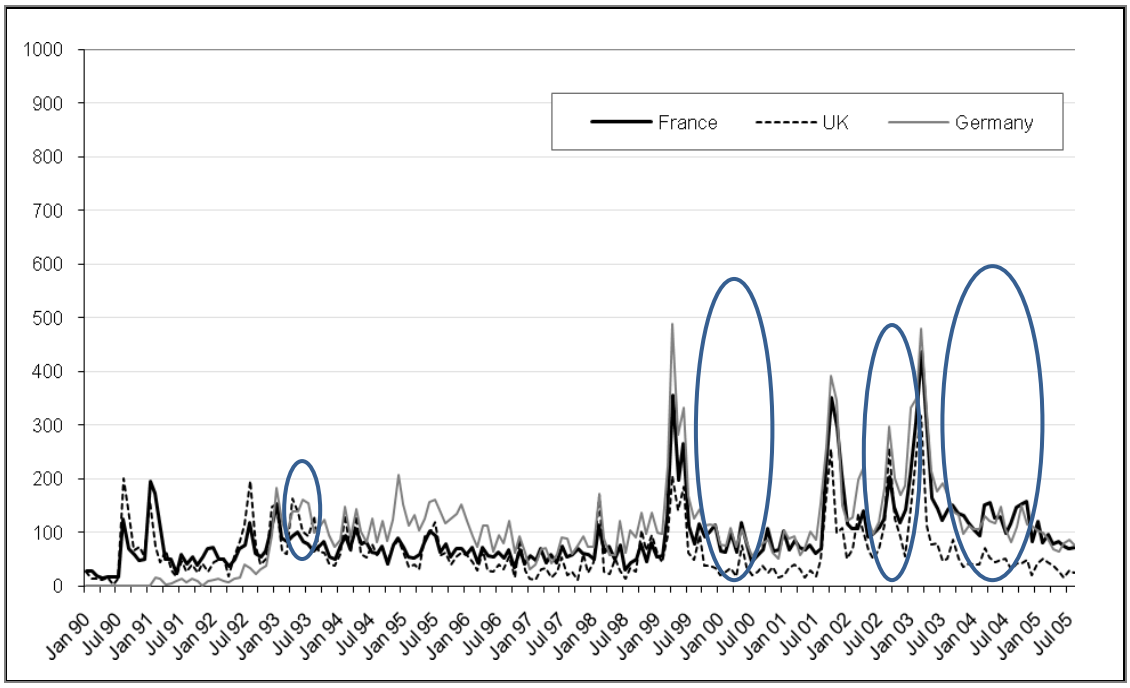
However, the very word ‘intervention’ (and its above explored semantic relations) was not particularly frequent in the corpus. This pointed to the abstract and specialist character of the term. A wordlist query in WordSmith revealed that commentators in the newspapers frequently used more palpable words that related to the field of military action and operation such as ‘troops’, ‘forces’, ‘strikes’, ‘attack’ etc. Therefore, we carried out an additional concordance analysis in WordSmith so as to isolate those collocation-node-spans that indicated ‘military and humanitarian’ in terms of the above definition. These were, for example, particular specifications of ‘troops’ (e.g. ‘UN troops’, ‘blue helmet’), ‘force’ (e.g. ‘monitoring force’), ‘forces’ (e.g. ‘contribute forces’, kfor, isaf), ‘missions’ (e.g. ‘military mission’, ‘peace-keeping’), ‘strikes’ (e.g. ‘NATO air strikes’). By means of this double-track procedure, we arrived at a consolidated list of term variants for each language that indicated the conceptual category ‘military and humanitarian intervention’. The results were cross-examined in language-contrasting team sessions to make sure that a semantic dimension that had proved more salient in one language than in another was nevertheless included in the general set of terms used for further quantitative assessment. Having thus identified a consolidated set of terms and node-collocate-spans that represented the conceptual category ‘intervention’, we could ‘mine’ the corpus for occurrences and distribution of the identified category.

Figure 5 illustrates one possible use of the category ‘intervention’ for content-analytical purposes. The figure displays the temporal distribution of those articles published in French, German, and British newspapers which contained one or several lexical realizations of the conceptual category ‘intervention’ (see Figure 5). The time series reveals that, in certain points of time, the salience of ‘intervention’ rises simultaneously and sharply in all the three countries (see blue circles in Figure 5). This is the case for: January 1993 when the Vance-Owen-Plan was presented that was meant to ensure a ceasefire between the armed groups in Bosnia; mid-1999 when the NATO ran air strikes against Belgrade in the attempt to stop the Kosovo conflict; autumn 2001 when the Al Qaida attack on the World Trade Centre in New York triggered a UN-sanctioned intervention in Afghanistan; summer 2002 when military action against Iraq was discussed in the UN security council; and spring 2003 when military action against Iraq was begun by a US-led coalition.

Along with the intersection of media coverage in moments of major conflict escalation and military action, Figure 5 also brings to light strong country-specific variation and changing tendencies of convergent agenda-setting. French and British newspapers raise the issue of

humanitarian and military intervention already in the early 1990s and follow it with similar attention throughout the following years. German newspapers, on the other hand, wake up to the problem only in relation to the Bosnian war, but after that cover the issue with greater intensity than do the newspapers in either France or the United Kingdom.

Figure 5: Articles Containing Lexical Realisations of 'Intervention' (Absolute Numbers)



Moreover, we can detect a shift from an alignment between British and French news coverage in the beginning of the 1990s to an alignment of French and German coverage from the Kosovo conflict on and, in particular, during the 2000s. The findings raise interesting questions about shifts in country-specific conceptions of foreign policy after the Cold War, in particular about the contestation and normalization of military intervention in formerly semi-dependent Germany. They also suggest that rapprochement in French and German press coverage of humanitarian and military intervention might be linked to perceptions that international conflict management is increasingly seen as a problem concerning the EU, both as political space and as international actor. The following sections assess this question in more detail.

4.2. Distinguishing the lexical field of 'EU-Europe'

To assess the perception of the macro region and potential political actor 'European Union', we sought to establish the EU's salience in news coverage on humanitarian and military intervention. The research question was thus: what news articles refer only to the European Union (its names, representatives, institutions, levels, procedures and policies), not to 'Europe' as a larger political or general geographic unit? When and how often and to what extent was the EU political space referred when international crisis events were mentioned, and hence portrayed as being involved in or affected by these events? The major problem was that everyday discourse and academic discourse alike do not employ separate wording, but refer to 'Europe' and 'European' regardless of whether only EU-Europe, the larger political Europe (e.g. including non-members) or the geographical unit is at stake. This was particularly true for the newspapers in English and French. Dutch, German-language, and (to some degree) Polish newspapers used the acronym 'EU' or compounds containing it to denote the EU political space, but not consistently. DISCO proved of limited help for disambiguating the semantic range of 'Europe': the measure of similar distribution retrieved information on other geographical units that were used in the same syntactical positions as 'Europe' in the corpus, but did not reveal the paradigmatic relations within the lexical field 'Europe'.

Consequently, we had to build the lexical field of 'EU-Europe' from introspection (judgement based on our knowledge of semantic clustering in the language) and detailed concordance analysis of collocates of the nodes 'Europe', 'European', 'Europeans', 'Union', and 'Brussels'. This analysis was informed by political science conceptions of EU-Europe as a political organisation and space that bears the characteristics of a 'negotiating multi-level system' (Kohler-Koch and Jachtenfuchs, 2004). The inductive analysis revealed that in all languages, but in French especially, media coverage on international conflicts employs a rich and fine-grained vocabulary to distinguish 'EU-Europe' from broader conceptions. Not only do commentators in the media employ specific EU wording and sophisticated paraphrases for particular institutions, policies, and objectives of European integration. They also differentiate between the various scales, centres and addressees of EU decision-making. In all languages and countries, words clustered in the following hyponymous dimensions of 'EU-Europe' that echo, to some extent, political science conceptions of the EU's multilevel system:

- EU names, acronyms (e.g. EEA), and paraphrases such as 'club européen', 'l'Europe a vingt';

- EU institutions ranging from EU decision-making, regulatory and juridical bodies and their representatives to legal institutions like EU treaties, EU citizenship, the Internal Market, or EU parliamentary elections;
- 'Brussels' as a territorial level ('European level', 'l'échelle européen') and power centre ('European compromise', 'accord de Bruxelles', 'eurocrat', 'directives de Bruxelles', 'pris à Bruxelles');
- intergovernmental politics signified by expressions like 'European governments', 'European leaders... ministers... ambassadors', 'capitales européennes', 'officiers européens';
- horizontal-transnational perspectives indicated by expressions like 'fellow Europeans', 'homologue européen', 'partners in Europe', 'nos frères européens' etc.
- national EU policies indicated by expressions like 'Britain and the rest of Europe', 'Britain's future in Europe', 'la France et l'Europe', 'appartenance a l'Europe', 'besoin de l'Europe';
- names and abbreviations of policies, bodies and representatives of EU-coordinated foreign, security, and defence policies and international aid;
- the foreign policy role and actorness of the EU and its members indicated by expressions like 'the EU should', 'the Europeans', 'entre l'Europe et…' etc. (see also section 4.3);
- visions and rhetoric of European unification indicated by expressions like 'stronger Europe', 'as Europeans, we', 'the Europe of', 'l'Europe unie' etc.

While all the languages and countries investigated showed some lexical realisations of these hyponymous dimensions, strong country-variation occurred in the degree of specification and elaboration. US newspapers were taciturn on most EU-specific dimensions, but differentiated in expressions signifying the security architecture and history of the larger (Cold-War) Europe, which had to be excluded from the lexical field 'EU-Europe'. And while all EU-based newspapers were more sophisticated in the wording of EU-Europe than the US newspapers, strong variation occurred between the countries in the specification of particular hyponymous dimensions of EU-Europe. French newspapers were especially detailed on paraphrases of EU names, institutional and horizontal dimensions of EU politics. British newspapers, to the contrary, were exhaustive on British EU policy, i.e. the measures taken by the British government or policies discussed by British politicians to address supranational decision-making from a British perspective. These subtle differences point to a more supranationalist perspective in French media discourse and a more intergovernmentalist perspective in British media discourse. They hint at country-specific conceptualisations of European integration.

These country-specific semantic preferences are particularly salient in the collocate-node-spans that lexicalise the dimension of Brussels-based supranational politics. While the British and the French newspapers shared a set of expressions for designating this dimension, the French had many more expressions than the British (see row 'Related to EU-Europe' in Table 1). Moreover, French newspapers articles on 'intervention' associated Brussels exclusively with the EU power centre. British newspapers, on the contrary, retained alternative word senses of 'Brussels' such as the NATO headquarters, the Brussels branch of the pan-European stock exchange Euronext, and the localities of the Belgian capital (see row 'Not related to EU-Europe' in Table 1).

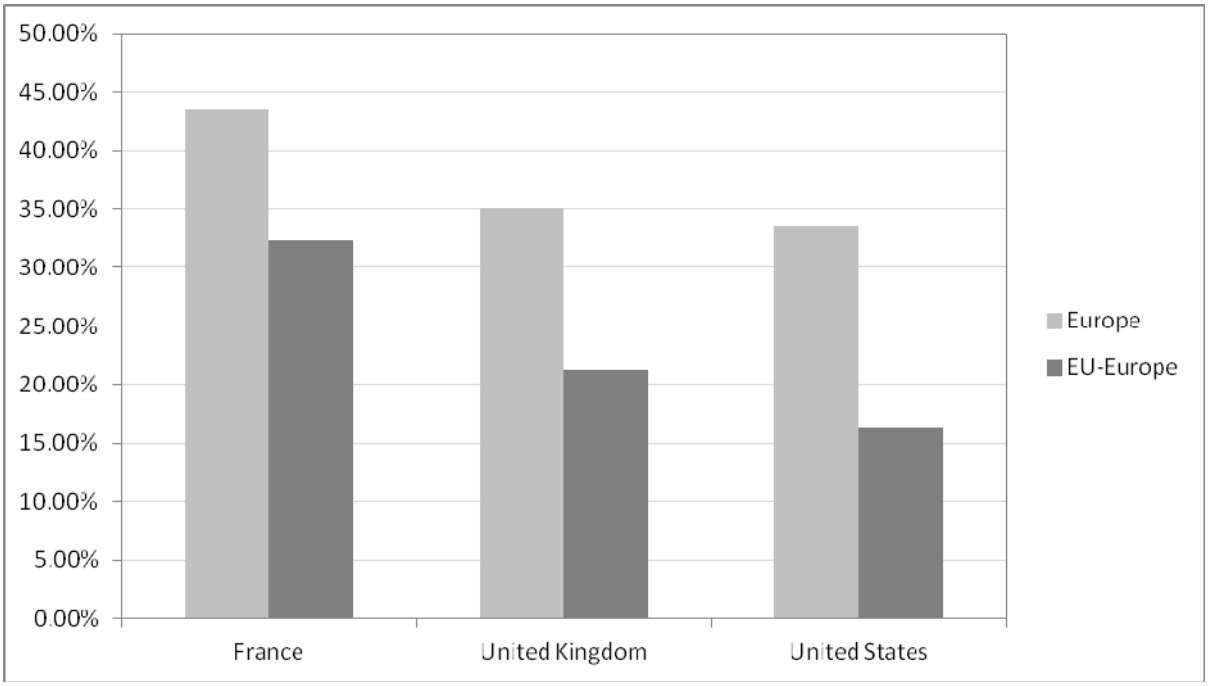
Table 1: Lexical Realizations of 'Brussels Politics' in Articles on 'Intervention'

	The Times, The Guardian	**Le Monde, Le Figaro, Les Echos**
Related to EU-Europe	Brussels bureaucrat Brussels bureaucracy Eurocrat European politics European agenda European summit Brussels official European level European policymaking European policy-making common European policy European consensus European framework	l'Europe de Bruxelles sommet européen calendrier européen l'agenda européen accord de Bruxelles compromis de Bruxelles décision de Bruxelles décisions de Bruxelles directive de Bruxelles directive européenne coup de pouce de Bruxelles fait remarquer Bruxelles décisions de Bruxelles règlement européen (and many more)
Not related to EU-Europe	from Brussels has in Brussels has Brussels' Euronext Brussels' headquarters Brussels' Chinese Brussels' Egmont Brussels' main square	_ _ _

The example of 'Brussels politics' demonstrates how different lexicalizations of EU-Europe may depend on the country and language in question. It exemplifies why a case-by-case study of concordances may be necessary for the composition of a set of lexical items which is unambiguously related to the investigated conceptual category (here: 'EU-Europe') in all the languages involved in the analysis. Once we had established the category 'EU-Europe' in its country- and language-specific realizations, we could take up the initial question and subject it to a comparative quantitative content analysis; namely the question of whether the EU-Europe was brought into connection with humanitarian and military intervention, and to what extent. Figure 6 shows the share of articles on 'intervention' that mention EU-Europe,

compared to the share of those that mention Europe as an unspecified larger political or geographical unit (see Figure 6). The figure reveals varying degrees of Europeanization of news coverage, i.e. varying degrees of salience of the region 'Europe' and of EU governance, more particularly. The French newspapers spoke much more often of 'Europe' *and* of 'EU-Europe' in relation to humanitarian and military intervention than their American and British counterparts. In fact, British and US newspapers seemed to have a roughly similar and lower concern for Europe as a region and continent. However, salience of 'EU-Europe' proved to be stronger in the United Kingdom than in the US (see Figure 6).

Figure 6: Articles Mentioning Europe / EU-Europe (in % of the Total of Articles on Intervention)



Note: the total of articles on intervention was: 12,097 (French newspapers); 22,899 (British newspapers); 36,151 (US newspapers).

These findings invite speculation about the reasons of variation, such as: established bilateral relations between the country, in which news coverage is analysed, and countries or regions in which conflict takes place; varying degrees of involvement in European politics and international peacekeeping etc. But they also suggest that an investigation might be useful as to whether more or less concern with the European macro-region and the political space of the EU corresponds to more or less concern with the role of the EU in international conflict management. Do commentators in the French newspapers, more than in British and American newspapers, conceive of the EU as a foreign policy actor who ought to act upon violent conflicts, or is the variation due to particular emphasis on conflict events where the

EU was involved? We sought to get a grasp on the answer by running a corpus-based content analysis of the conceptual category 'EU actorness'.

4.3. Identifying semantic indicators of 'EU actorness'

In political science, 'international actorness' refers to the capacity of an international organisation or state to act intentionally in relation to other actors in the international system (Groenleer and Van Schaik, 2007, Sjöstedt, 1977, Smith, 2003). Jupille and Caporaso (1998) identified four dimensions that specify such action capacity: cohesion, authority, autonomy and recognition. *Cohesion* refers to the degree to which an entity, (in this case, the EU) is able to formulate and articulate internally consistent policy preferences. *Authority* pertains to the legal competence of the EU to act. *Autonomy* implies institutional distinctiveness, meaning that the EU can operate relatively independently from individual EU Member States. *Recognition* refers to acceptance of and interaction with the EU by others (Groenleer and Van Schaik, 2007).

Researchers agree that, from the perspective on institutional capacities and with regard to most of the above dimensions, the EU only partially qualifies as international actor, even though the recent treaty revision (Lisbon Treaty) endowed the EU with an integrated legal personality, a foreign minister and a diplomatic service. Still, this does not preclude that the EU is perceived as acting, and demanded to act, as if it were an international actor. Bengtsson and Elgström point out that such role constructions exist, both in self-conceptions of the EU as 'normative great power' (emphasising a civilising mission, a normative agenda, and civilian measures) and in role prescriptions of EU outsiders, who either affirm or obstruct this view (Bengtsson and Elgström, 2009). These role attributions may be seen as a form of recognition of the EU's actorness. Even if a partner denied the EU the capacity to act in international politics, this would still conceptualise the EU in terms of actorness, measuring its performance according to criteria of cohesion, authority, and autonomy that other political organisations are expected to have. Hence, the very fact that the EU is seen as having or missing actorness might enhance recognition.

Following this consideration, we sought to find out whether portrayals of the EU as international actor (or as an organisation missing these qualities) were also salient in media coverage and debate on intervention. If this were the case in several EU member states with regard to the same conflict events, then this could indicate a shared concern with the EU's international role and concrete action potential in conflict management and, hence, an emerging EU-related problem-solving community (Kantner et al., 2008). In line with this

expectation, we sought to establish whether, with regard to which conflicts and with what intensity the EU was portrayed as an entity capable of acting, obliged to act, or failing to act upon international conflicts in the news articles of our corpus. However, even more than 'intervention' or 'EU-Europe', 'actorness' was unlikely to have any straight-forward equivalent in media language. Indeed, the terms 'actor', 'actorness', 'action capacity' etc. did not at all occur in the corpus. Therefore, we had to identify, through concordance analysis, the syntagmatic relations that constituted a lexical realm close to the notion of actorness.

Table 2: Lexical Realizations of 'EU Actorness' in Articles on 'Intervention' (Extracts)

	The Times, The Guardian	**Le Monde, Le Figaro, Les Echos**
Prepositions and conjunctions of relation	between Europe Europe's relations with relations with Europe ties with Europe between the United States and Europe...	entre l'Europe et du côté de l'Europe vis-à-vis de l'Europe rélations avec l'Europe rélations Europe - sommet Europe- les Etats-Unis et l'Europe...
Possesive constructions and attributions	European initiative European position Europe's capacity Europe's failure Europe's role European interests European responsibility time for Europe...	initiatives européennes position européenne capacité européenne puissance de l'Europe faiblesse de l'Europe rôle de l'Europe responsabilité de l'Europe heure de l'Europe...
Subject-predicate and modal constructions	if Europe that Europe is that Europe had that Europe faces . Europe will . Europe should . Europe would . Europe could . Europe may . Europe needs . Europe has . Europe is . Europe cannot . Europe ought . Brussels must...	l'Europe doit l'Europe ne doit l'Europe devrait l'Europe ne devrait l'Europe doive l'Europe ne doive l'Europe peut l'Europe ne peut . L'Europe n'est . L'Europe n'a . L'Europe s'est . L'Europe sera . L'Europe reste . L'Europe fait . Bruxelles devrait...

A detailed assessment of collocates of the node 'Europe' and 'European' revealed several instances of colligation and semantic preference that were unambiguously related to EU actorness in all the languages investigated. These were prepositional clauses and conjunctions that juxtaposed 'Europe' vis-à-vis other international actors (e.g. 'between Europe and the US'; 'Europe, Japan, and...'); possessive constructions that attributed terms from the lexical field of strategic (foreign policy) action to the EU (e.g. 'European positions',

'Europe's failure', 'time for Europe'); and prepositional clauses, modal constructions, or particular predicative clauses that identified 'Europe' as subject acting upon something (e.g. 'Europe should...'; 'Europe has...'). Table 2 reveals surprising parallels in the way the international actor EU is lexicalised in British and the French newspapers (see Table 2). The table also shows that those node-collocate-spans that construct the European Union as grammatical subject (the EU should...) are often in modal categories that express probability, possibility and necessity. The semantic preference between the grammatical subject EU and modality points to an only provisional recognition of the EU's actorness; it indicates that the EU is mainly portrayed as would-be actor of international conflict management.

The so-grounded concept of 'EU actorness' was subjected to a quantitative analysis and assessed with regard to its salience and distribution throughout the corpus. The analysis revealed that, during the period of investigation, portrayals of the EU as playing or not playing a role in humanitarian and military interventions were not particularly frequent. The share of articles on 'intervention' that contained lexical realisations of 'EU actorness' ranged between 19.6 percent in the French newspapers, followed by Austrian newspapers (18.4 percent); 16.1 percent in Irish newspapers; 15.5 percent in the German newspapers, followed by Dutch newspapers (13.4 percent); and 10.4 percent in British newspapers, whose low score was only beaten by the US newspapers (6.2 percent). Hence, the findings on the salience of 'EU-Europe' presented in the previous section were repeated for the hyponym of 'EU-Europe': the category 'EU actorness'. On the one pole of the continuum were French newspapers that stood out by frequent reference to the EU and its actorsness in relation to humanitarian and military intervention (here followed by some small member states with a tradition in neutrality and Germany); on the other pole were American newspapers that paid little attention to 'EU actorness'. And UK newspapers, again, seemed to take a middle position.

Regardless of varying degrees of media salience of the EU's international role, commentators in all the EU-based newspapers agreed on the conflict events in relation to which portraying the EU as an actor made sense. This was revealed by an investigation of the co-occurrences, within an article, of the category 'EU actorness' and names of those regions or countries that had witnessed violent conflicts during the period of investigation (1990-2006). Overwhelmingly, lexical realizations of 'EU actorness' co-occurred with the Western Balkans (a label here comprising former regions and succesor states of Yugoslavia), in particular in the beginning and the end of the 1990s. Second came co-occurences with the Middle East (a label here comprising Israel, Syria, Lebanon, Palestine, Iran, Iraq, Kuwait, Yemen), in particular in the years of the US-led intervention in Iraq. Hence,

from the hot-spots of coverage on 'intervention' identified in section 4.1 – the Western Balkans and the Middle East – the EU is more likely to emerge as potential actor. Having said this, the comparison of the temporal distribution of 'EU actorness' in French, British, and German newspapers also revealed that the EU's international role lost salience in British newspapers from 2002 on. On the contrary, French and German newspapers displayed an increased attention for 'EU actorness' after mid-2002.

This points to a response to the question posed at the end of previous section. As expected, national differences in the salience of 'EU-actorness' do coincide with differences in the salience of 'EU-Europe'. This reinforces the hypothesis that, once the political space of 'EU-Europe' is brought into connection with humanitarian and miliary intervention, portrayals of the EU as international actor are more likely to emerge. But national variation in salience is not, apparently, due to varying emphasis on different conflict events. In fact, when the EU is lexicalised as actor, it is predominantly in relation to the same conflict events. The divergent development between the British vs. the French and the German newspapers regarding 'EU actorness' in the 2000s rather points to the intra-European conflict on the Iraq war, with French and German public opinion closing ranks in emphasising a 'European' perspective as opposed to the approach taken by the US administration and its allies. But above all, the divergence in emphasis on the EU as a would-be actor is likely to result from varying attention for the EU's foreign, security, and defense policies which took concrete shape with the negotiation of the European Security Strategy (adopted in 2003) and the Constitutional Treaty (adopted in 2004, now Lisbon Treaty). With the progressive institutionalisation of joint foreign and security policies, the EU bodies, in fact, started acting or presenting themselves as if they were or ought to be a single international actor. Most interestingly, though, this is not necessarily reflected in media coverage. The British example shows that public recognition of the EU as an international actor is not given with shifts in competences. Rather, it depends on a broader recognition of the relevance of the political space 'EU-Europe'.

5. Conclusions

This paper sought to clarify how quantitative content analyses of large text samples can be conducted so that inference-making is both representative of the actual meaning of the analysed text and indicative of the conceptual category that the researcher seeks to investigate. The challenge of valid and reliable inference-making is, in part, implied in the content analytical method itself. Content analysis derives categories from social theory, not

from the actual make-up of the text. It pertains to answering questions about the socially relevant mental and interpretative concepts, but not about semantic coherence and syntactic cohesion of a text, and respective linguistic theorising. Hence, there is always a gap between the 'real object' of the text and the 'inferred object'. This becomes a problem when particular parts of an analysed texts are no longer classified in light of the entire text item (as it would be the case with a qualitative content analysis of small samples), but classified automatically. Here, misrepresentation of the text is very likely; and the need for an intermediary, smaller ranged, theory and a text-sensitive quantitative method becomes obvious.

The authors suggested using semantic field theory and corpus semantic analytical procedures as a means for the mediation between conceptual categories of the social sciences on the one hand, and the lexis of the concrete texts compiled for primary analysis on the other. Semantic field theory assists in viewing conceptual categories from a semantic perspective – as constituted through key terms and semantically related terms. It suggests building and disentangling lexical clusters according to paradigmatic (e.g. synonyms, hyponyms) and syntagmatic (sequence-related) relations between words.

Corpus semantic analytical procedures, whether based on manual or automated assessment of lexical-syntactic characteristics of words in a given text collection, facilitate the systematic disambiguation of word senses and the identification of relevant semantic clusters. In our project, they helped to inductively detect those term variants and word senses that were indicative of broad concepts like 'military and humanitarian interventions', 'EU-Europe' and 'actorness'.

Based on the insights from semantic field theory and corpus semantics, we turned the conventional content analytical procedure 'corpus-based'. We adapted it in a way which ensured sensitivity towards the actual text and the natural languages involved (through inductive-manual corpus analysis) and prepared the quantitative analysis of the salience and interrelation of these concepts (through automated text mining of the corpus). *Corpus-based content analysis* ideally includes the following steps:

- Concept specification: specification of key terms and associated terms (rather than content dimensions) that are central to the investigated conceptual category, based on relevant theoretical literature;
- Concept identification: language-contrasting investigation of lexical approximations of these terms (rather than content variants) in the corpus under investigation with the help of NPL based tools and corpus-linguistic wordlist and concordance analyses, creation of

list of search terms or node-collocate-spans that unambiguously evoke the investigated conceptual category in all languages analysed;

- Concept assignment: text mining of occurrences of items from the search word list (instead of hermeneutic-interpretative coding);
- Quantitative analysis: descriptive or analytical statistic assessment of frequency, co-occurrence or correlation of 'mined' conceptual categories.

The examples from the investigation of news coverage on war and intervention showed that such procedure yields valuable insights in two respects. The comprehensive *inductive corpus analysis* brought to light the rich vocabulary of international conflict management and European politics used in news coverage on international conflicts. With regard to EU politics, in particular, the analysis revealed pronounced country-specific (rather than language-specific) lexicalizations. They pointed to the predominance of certain conceptualisations of European integration in the national media analysed, e.g. more stress on the supranational dimension in French newspapers as opposed to more stress on the intergovernmental dimension in the British newspapers. The *descriptive-statistical analysis* of occurrences and distributions of lexicalizations of the conceptual categories produced an overview of the trajectory of the agenda, selected issues and perceptions in news coverage on war and military intervention. It revealed tendencies of convergence or divergence between the different national media over time. These quantitative data, in particular when placed in relation to each other, allowed for the investigation and further qualification of the hypotheses from which the project had started, i.e. hypotheses on the emergence of problem-solving communities through transnationally convergent issue-related debate and problem assessment.

For instance, the query of co-occurring lexicalizations 'EU actorness' and names of conflict regions in articles on 'intervention' in section 4.3 revealed that problem assessments converged with regard to *when*, in relation to what conflicts, the EU was seen as obliged or (in-)capable of engaging in conflict management. Yet, the degree to which the EU was portrayed as such varied strongly between countries. Hence, as illustrated by Figure 7, the mining of the occurrences of (lexicalizations of the) different conceptual categories in the corpus not only produced several sub-samples that could be assessed in more detail, but these sub-samples could also be used as content analytical 'variables' to be further investigated by means of analytical statistics (as in Kantner, 2009).

Figure 7: Sub-Samples or: Variables of Content Analysis

Naturally, the proposed methodology of *corpus-based content analysis* has its particular limitations. Due to the focus on semantic fields of key terms, we only assessed the salience of semantic-lexical characteristics of single words or word groups. More complex, sentence- or text-based, constructions were not considered.[14] As a result, we could 'mine' only lexical-thematic aspects of news coverage. However, as the investigation of 'EU actorness' showed, the analysis can be extended to evaluative aspects when focussing on word-related discourse prosodies that imply evaluation. The fact that mentions of the EU as an internationally acting subject were mostly put forth in modal constructions (the EU should, could, must etc.) suggests that EU actorness is primarily seen as (unrealised) possibility and obligation. And frequent mention of 'European failures' or 'incapacities' in relation to humanitarian and military intervention implied clear-cut, negative, discourse prosody. Here, thematic analysis borders on phenomena that, in conventional content analysis, are understood as 'framing' or 'priming'.

[14] Advanced NPL-based methods consider word, sentence, and text levels (Kolb, 2008, Mehler et al., 2011, Stede, 2008, Wüest et al., 2011).

Another, related, constraint is the simple application of corpus linguistics and of natural language processing. The corpus linguistic part of the analysis was limited to the scrutiny of salient key words and alternative concordances lexicalising the investigated conceptual category. And instead of sophisticated linguistic annotations, we 'mined' simple character strings that we had constructed manually while doing the inductive corpus analysis. However, little linguistic and computational sophistication means that the procedure can be easily reproduced by social scientists who do not have advanced knowledge of linguistics or programming skills, provided they are willing to familiarise themselves with the basics of semantics. More importantly, the manual construction of lexical items from meticulous concordance analyses ensured high internal validity: only those lexical items were included in text mining and quantitative analysis that, during the corpus analysis, were classified as unambiguously embodying the conceptual category. It is the 'manuality' that warrants both the grounding of conceptual categories in the linguistic structure of specific natural languages and the abstraction that is implied in content analysis. Through manual corpus analysis, we retained the hermeneutic moment of content analysis: we inferred directly from the concordances to the conceptual categories. At the same time, it was with the help of corpus linguistic procedures (word displays reduced to node-collocate-spans) and computational linguistic tools from distributional semantics that we could do this in an effective manner, for huge samples of texts.

It remains to conclude with a general precaution: despite its simplicity and semi-automation, the proposed corpus-based content analysis is labour intensive. This is not only due to the detailed corpus analysis involved. Any quantitative content analysis implies much more than word counting (Roberts, 1989). The recommended procedure is labour intensive in particular because of the complex methodological and technical issues that come with the compilation and storage of digital newspaper corpora and their pre-processing for corpus analysis and text mining (Gabrielatos, 2007, Kantner et al., 2011). For this reason, corpus-based content analysis is preferably applied to 'sustainably built' text samples which can be queried for various research questions in the field. The corpora should also be large enough to yield sound frequency data in corpus and computational linguistic tools.[15] Having said this, we believe that we have proposed a method that addresses some of the major challenges of content analysis which have surfaced with the availability of massive samples of multilingual digital texts.

[15] Estimations of what minimum number of tokens is needed to yield reliable rather than arbitrary patterns in corpus computation range between 300,000 and one million words. The experience value in the project was ca. 60,000 news articles of medium length. Beyond this threshold, WordSmith tracked a critical mass of regular collocates, patterns, and clusters that allowed judgement on the lexical field related to a key term of a conceptual category.

6. References

Baker, P. (2006) Using Corpora in Discourse Analysis. London: Continuum.

Baker, P., Gabrielatos, C., Khosravinik, M., et al. (2008) A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press, Discourse & Society, 19(3), 273-305.

Baker, P. and McEnery, T. (2005) A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts, Journal of Language and Politics, 4(2), 197-226.

Bengtsson, R. and Elgström, O. (2009) The European Union's International Roles. 5th General ECPR Conference. University Potsdam.

Fühlau, I. (1982) Die Sprachlosigkeit der Inhaltsanalyse. Linguistische Bemerkungen zu einer sozialwissenschaftlichen Analyse. Tübingen: Gunter Narr Verlag.

Gabrielatos, C. (2007) Selecting query terms to build a specialised corpus from a restricted-access database, ICAME Journal, 31, 5-43. (http://www.gabrielatos.com/RQTR.htm, Access date: 02-21- 2012)

Gliozzo, A. (2006) Semantic Domains and Linguistic Theory. Language Resources and Evaluation Conference - LREC. Genoa, 5.

Gliozzo, A. and Strapparava, C. (2009) Semantic domains in computational linguistics. Berlin [u.a.]: Springer, 15-16.

Grabowsky, J.-K. (2011) Who Cares about Genocide in Europe? Identity-Related Reactions to Interventions and the Srebrenica Massacre in French, German, Dutch and American Newspapers. Otto-Suhr Institute of Political Science. Berlin: Free University Berlin.

Grimmer, J. and King, G. (2011) General Purpose Computer-Assisted Clusterin and Conceptualization, Proceedings of the National Academy of Sciences108, 108(7), 2643-2650.

Groenleer, M. L. P. and Van Schaik, L. G. (2007) United we stand? The European Union's international actorness in the cases of the International Criminal Court and the Kyoto Protocol, Journal of Common Market Studies, 45(5), 969-998.

Holsti, O. R. (1969) Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley.

Jupille, J. and Caporaso, J. A. (1998) States, agency and rules: The European Union in global environmental politics. In: C. Rhodes (ed) The European Union in the World Community. Boulder, CO: Lynne Rienner, 213-229.

Kantner, C. (2009) Transnational Identity-Discourse in the Mass Media. Humanitarian Military Interventions and the Emergence of a European Identity (1990-2006). Habilitation thesis. Berlin: Freie Universität.

Kantner, C. and Kutter, A. (2011) International Crisis as Engine of Institutionalization? Media Debates on Armed Conflicts and Development of the European Security and Defense Policy (1990-2005). ISA Annual Convention. Montreal, Kanada.

Kantner, C., Kutter, A., Hildebrandt, A., et al. (2011) How to Get Rid of the Noise in the Corpus: Cleaning Large Samples of Digital Newspaper Texts [in collaboration with: Andreas Hildebrandt and Mark Püttcher], International Relations Online Working Paper (2011/02). (http://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/IROWP_Series_2011_2_Kantner_Kutter_Analysis_Newspaper_Texts.pdf, Access date: 02- 21- 2012)

Kantner, C., Kutter, A. and Renfordt, S. (2008) The Perception of the EU as an Emerging Security Actor in Media Debates on Humanitarian and Military Interventions (1990-2006), RECON Online Working Paper, 2008(19)**,** Arena Oslo, Oslo. (http://www.reconproject.eu/main.php/RECON_wp_0819.pdf?fileitem=16662569, Access date: 02- 21- 2012)

Kantner, C. and Renfordt, S. (2007) Public Debates on Humanitarian and Military Interventions in Europe and the U.S. (1990-2005). Paper presented at the 48th Annual ISA Convention 'Politics, Policy and Responsible Scholarship', Chicago, IL Feb. 28th - March 3rd, 2007. Chicago. (http://citation.allacademic.com/meta/p_mla_apa_research_citation/1/8/0/5/3/pages180532/p180532-1.php, Access date: 02- 21- 2012)

Kennedy, G. (1998) An Introduction to Corpus Lingustics. London, New York: Longman, 247-256.

Kilgariff, A. (2003) Thesauruses for Natural Language Processing. Proceedings of Natural Language Processing and Language Engineering. Bejing.

Knapp, W. (2008) Die Inhaltsanalyse aus linguistischer Sicht. In: P. Mayring and M. Gläser-Zikuda (eds) Die Praxis der Qualitativen Inhaltsanalyse. 2nd ed. Wineheim; Basel: Beltz, 20-36.

Koenig, T., Mihelj, S., Downey, J., et al. (2006) Media framings of the issue of Turkish accession to the EU, Innovation: The European Journal of Social Sciences, 19(2)**,** 149-169.

Kohler-Koch, B. and Jachtenfuchs, M. (2004) Multi-Level Governance. In: A. Wiener and T. Diez (eds) European Integration Theory. Oxford: Oxford University Press.

Kolb, P. (2008) DISCO: A Multilingual Database of Distributionally Similar Words. In: A. Storrer, A. Geyken, A. Siebert, et al. (eds) KONVENS 2008 - Ergänzungsband: Textressourcen und lexikalisches Wissen. Berlin.

Krippendorff, K. (2004) Content Analysis: An Introduction to Its Methodology. London: Sage.

Kutter, A. (2011) Polity-Construction in Multilevel Settings. Recontextualisation and the Exampled of the Polish and French Media Debates on the EU constitution. Department of Cultural Studies. Frankfurt/O.: European University Viadrina.

Lenz, S. (2000) Korpuslinguistik. Tübingen: Julius Groos Verlag Brigitte Narr.

Liebert, U. (2007) Introduction: Structuring Political Conflict about Europe: National Media in Transnational Discourse Analysis, Perspectives on European Politics and Society, 8(3), 236-260.

Magnini, B., Strapparava, C., Pezzulo, G., et al. (2002) The role of domain information in word sense disambiguation, Natural Language Engineering, 8(4), 359-373.

Mayring, P. (2002) Qualitative Content Analysis: Research Instrument or Mode of Interpretation? In: M. Kiegelmann (ed) The role of the researcher in qualitative psychology. Tübingen: Verlag Ingeborg Huber, 139-148.

Mehler, A., Diewald, N., Waltinger, U., et al. (2011) Evolution of Romance Language in Written Communication: Network Analysis of Late Latin}and Early Romance Corpora, Leonardo, 44(3).

Renfordt, S. (2010) How International Law Standards Pervade Discourse on the Use of Armed Force: Insights into European and US Newspaper Debates between 1990 and 2005, KFG Working Paper (2010/13). Freie Universität, Berlin. (http://www.polsoz.fu-berlin.de/en/v/transformeurope/publications/working_paper/index.html, Access date: 02- 21-2012)

Renfordt, S. (2011) Framing the Use of Force: An International Rule of Law in Media Reporting. A Comparative Analysis of Western Debates about Military Interventions, 1990-2005. Baden-Baden: Nomos.

Roberts, C. W. (1989) Other than counting words: A linguistic approach to content analysis, Social Forces, 68(1), 147-177.

Simon-Vandenbergen, A.-M. and Aijmer, K. (2002) The expectation marker of course in a cross-linguistic perspective, Languages in Contrast, 4, 13-43.

Sjöstedt, G. (1977) The External Role of the European Community. Westmead: Saxon House.

Smith, K. E. (2003) European Union Foreign Policy in a Changing World. Cambridge, MA: Polity Press.

Stede, M. (2008) Local Coherence Analysis in a Multi-Level Approach to Automatic Text Analysis, Journal of Language Technology and Computational Linguistics, 23(2), 1-19.

Stone, P. J. (1966) A perspective on content analysis. In: P. J. Stone, D. Dunphy, M. S. Smith, et al. (eds) General Inquirer: A computer approach to content analysis. Cambridge, MA: MIT Press, 3-19.

Stubbs, M. (2001) Words and Phrases. Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Tognini-Bonelli, E. (2001) Corpus Linguistics at Work. Amsterdam, Philadelphia: John Benjamins.

Trenz, H.-J. (2004) Media coverage on European governance. Exploring the European public sphere in national quality newspapers, European Journal of Communication, 19(3)**,** 291-319.

Trier, J. (1931) Der deutsche Wortschatz im Sinnbezirk des Verstandes. Heidelberg: Winter.

Ullmann, S. (1951) The Principles of Semantics. Glasgow: Jackson, Son & Co.

Vassilyev, L. M. (1974) The theory of semantic fields: a survey, Linguistics, 137**,** 79-93.

Weber, R. P. (1990) Basic content analysis. Newbury Park: Sage.

Wüest, B., Clematide, S., Bünzli, A., et al. (2011) Semi-Automated Core Sentence Analysis: Improving Content Analysis for Electoral Campaign Research, International Relations Online Working Paper (2011/01). (http://www.uni-stuttgart.de/soz/ib/forschung/IRWorkingPapers/IROWP_Series_2011_1_Wueest_Clematide_Buenzli_Laupper_Content_Analysis.pdf, Access date: 02- 21- 2012)