Pragmatics 21:4.647-683 (2011)
International Pragmatics Association

DETECTING CONTRAST PATTERNS IN NEWSPAPER ARTICLES BY COMBINING DISCOURSE ANALYSIS AND TEXT MINING

Senja Pollak, Roel Coesemans, Walter Daelemans, and Nada Lavrač

Abstract

Text mining aims at constructing classification models and finding interesting patterns in large text collections. This paper investigates the utility of applying these techniques to media analysis, more specifically to support discourse analysis of news reports about the 2007 Kenyan elections and post-election crisis in local (Kenyan) and Western (British and US) newspapers. It illustrates how text mining methods can assist discourse analysis by finding contrast patterns which provide evidence for ideological differences between local and international press coverage. Our experiments indicate that most significant differences pertain to the interpretive frame of the news events: whereas the newspapers from the UK and the US focus on ethnicity in their coverage, the Kenyan press concentrates on sociopolitical aspects.

Keywords: Text mining; Discourse analysis; Pragmatics; Ideology; Kenyan elections.

1. Introduction

Knowledge discovery in databases (Fayyad et al. 1996) is an iterative process of searching for valuable information in large volumes of data in a cooperative effort of humans and computers: humans select the data to be explored, define analysis problems, set goals and interpret the results, while computers search through the data, looking for models and patterns that meet the human-defined goals. The central step in this process is *data mining* (Witten and Frank 2005), the purpose of which is to automatically build classification models or find descriptive patterns in large data collections. A variant of data mining is *text mining* (Feldman and Sanger 2007) where models and patterns are extracted from collections of text documents. Text mining is relevant to linguistic research thanks to its ability to (i) process large amounts of text, which is hard to do by hand, and (ii) automatically uncover non-obvious and unexpected patterns in language use, for example in newspaper discourse. The goal of this paper is to investigate the potential of text mining techniques for *pragmatic discourse analysis*, where the purpose of *pragmatics* is to investigate how language functions in concrete socio-cultural contexts as a complex form of social action by looking for patterns of (explicit and implicit) meanings, dynamically generated in the process of using language (Verschueren 1999). The combination of text mining and

discourse analysis from a pragmatic perspective, has – to the best of our knowledge – not yet been explored.

In the case of knowledge discovery using text mining, the discovery process consists of the following steps: selection of a corpus, document preprocessing, text mining, and finally the interpretation and human evaluation of the automatically discovered models and patterns. In our study, the latter two steps are performed from a linguistic pragmatic perspective.

Our paper has both a methodological and a thematic aim. Methodologically, this paper aims to investigate whether text mining techniques can be applied in support of pragmatic news discourse analysis and so illustrate the fruitful interaction of these two approaches. Our second aim is to scrutinize the language use in newspaper reports about the December 2007 Kenyan elections and the ensuing post-election crisis, taken from the Kenyan newspapers *Daily Nation* and *The Standard* as opposed to the British and American newspapers *The Independent*, *The Times*, *The New York Times* and *The Washington Post.*

The starting hypothesis is that a comparison of different newspaper articles will show a discrepancy between local (Kenyan) and international ('Western') news coverage, which can partly be accounted for in terms of ideology. To test this hypothesis, the results of text mining will be interpreted and evaluated from the perspective of linguistic pragmatics, which studies cognitive, social and cultural aspects of language use. Language is seen as a process of meaning generation characterized by the constant making of choices, at both production and interpretation levels, from a variable and varying range of structural and contextual options. These choices are made in a flexible, negotiable manner and inter-adapt to reach relative satisfaction for communicative needs (Verschueren 1999, 2008). According to this pragmatic framework, all (conscious and unconscious) lexical, syntactic or discursive choices, relating to the vocabulary used, syntactic structures, modality, information structure, textual organization, etc., are considered to be significant and could have ideological implications. In this research, we mainly focus on the words used as well as salient absences of specific (often value-laden) lexical choices.

This paper is structured as follows. In Section 2 we describe the related work in computer assisted discourse studies and previous studies of text mining applied to media analysis. Section 3 provides the motivation for this work by discussing the role of ideology in news reporting. In Section 4 the studied corpus of newspaper articles is presented in its historical and political context, together with the initial corpus exploration using a standard corpus linguistics tool. In order to discover interesting differences between local and Western media, we used several data mining and text mining tools leading to results presented in Sections 5 and 6. Detailed analysis of the results from a pragmatic perspective is presented in Section 7. The paper concludes with the summary of main differences in news reporting and with our analysis of the complementarities of the two methodological backgrounds, text mining and pragmatics-based discourse analysis.

2. Related work

Computer technology is useful for the analysis of discourse and different techniques have been used for many years in discourse studies. However, computer tools are not

yet part of the mainstream methodology, especially not in 'handcraft disciplines' like critical discourse analysis (CDA) or socio-culturally oriented linguistic pragmatics. While an exhaustive survey of computer-assisted discourse studies with comparisons of the different methods falls outside the scope of this paper, we present the major strands of related work without going into details.

The related work first addresses the main computer-assisted approaches that can be helpful for pragmatists or discourse analysts: Computer-assisted qualitative data analysis software and corpus linguistics. This is followed by a brief review of related work in text mining applied to media analysis. We describe and position our approach in relation to these research areas.

2.1. *Computer-assisted discourse studies*

Computer-assisted qualitative data analysis software (CAQDAS), a term coined by Fielding and Lee (1998), broadly refers to software "that supports a variety of styles in qualitative research" (Gibbs 2004: 87). CAQDAS is available through software packages, typically incorporating functions for text searching, coding, transcription and data visualization (Lindlof and Taylor 2011: 260–266). Examples of software packages are NVivo, MAXQDA and Atlas.ti. These tools, which can be used for quantitative linguistic analysis, have a greater overlap with corpus linguistics tools (MacMillan 2005: 17–22, see also below) than with text mining tools. In fact, in discourse analysis they are primarily used as tools for data management. Although they are useful for searching through collections of texts, for (interlinked) coding and for retrieving (coded) data segments, some discourse analysts find them time-consuming, impractical and distracting from both their research focus and the data, or even claim that they could easily lead to senseless or random, hyperactive coding and may invite the user to draw unwarranted conclusions (e.g. MacMillan 2005; Schönfelder 2011).

Linguistic corpora (Sinclair 1991) and *corpus linguistics* (Kennedy 1998) have a long research tradition. Corpus linguistics can be generally defined as "the study of 'real life' language use with the help of computers and electronic corpora" (Lüdeling and Kytö 2008: v), or as "the study of machine-readable spoken and written language samples that have been assembled in a principled way for the purpose of linguistic research" (O'Keeffe et al. 2011: 6). As such, it lends itself perfectly to support discursive or pragmatic analysis of language in use, especially when a large amount of text is involved (cf. Koller and Mautner 2004). A corpus linguistic approach to discourse studies provides quantitative evidence of the existence of certain (ideological) discourses by enabling researchers to identify repetitive linguistic patterns of language use and to uncover hidden meanings, particularly in lexical items, e.g. by examining collocations (Baker 2006). Allowing some generalization, there are two basic kinds of 'corpus-driven' discourse studies (see Stubbs 1996; O'Halloran 2010; Thornbury 2010 for overviews). In the first kind, the use and the meanings of an *a priori* chosen linguistic expression or part of speech is studied in a language or in a specific discursive domain by looking for frequencies, co-texts and collocations, whereby the quantitative results are linked to society and are given social interpretations (e.g. *ethnic, racial* and *tribal* in Krishnamurty 1996, *elderly* in Mautner 2007 or *sustainable development* in Mahlberg 2007). In the second kind of research,

corpus linguistic methods are used in function of further critical discourse analysis: meaning patterns can be revealed by identifying and quantifying emerging keywords or collocations so as to create an insight into a specific discourse and boost or guide detailed and contextualized analysis of the discourse (e.g. Baker et al. 2008; Morley and Bayley 2009). Another way is to use methods from corpus linguistics, not at the start, but during or after a pragmatic or critical analysis of language use, as a kind of triangulation to support and reinforce the interpretive results of the discourse analysis (e.g. O'Halloran and Coffin 2004; Baker 2006; Baker et al. 2008).

Our approach shows some similarity with this work. For one, our endeavor is inspired by the same basic philosophy of combining computer-assisted quantitative analysis with in-depth qualitative analysis to improve 'manual' analysis of discourse and give the interpretive results more rigor (e.g. Koller and Mautner 2004; O'Halloran 2010). Furthermore, we share a focus on meanings and real language use in context. However, we have different conceptions of meaning and context. In addition, we believe that our methodology combining discourse analysis and text mining can open up new pathways into discourse understanding.

Most differences are due to the computational methods that are used. Corpus linguists typically use frequency, keyword or concordance-based pattern analysis techniques, while data mining methods, originally used for predictive purposes, do not concentrate only on frequent words but on the most distinctive words and their combinations which uncover patterns characterizing groups of documents: when two classes of documents are studied, such as local and Western press articles, our approach is well suited for detecting contrast patterns between these document groups.

Since one of the main drawbacks of text mining is that the words are taken out of the context, we creatively exploit text mining tools in combination with linguistic pragmatic analysis to get a deeper insight into naturally occurring, contextualized discourse. The meanings corpus linguists are concerned with are often limited to semantic meanings of individual lexical items, while we are interested in the meanings of the discourse as a whole as well as the pragmatic functioning of language in use. Also the notion of context is more narrow: "In corpus work, context means […] not only co-text (a short span of a few words within one single text), but also inter-text (repeated occurrences, often a very large number, of similar patterns across different, independent texts)" (Stubbs 2001: 157). In our pragmatic framework, context is broadly conceived as "[a]ny (combination of) ingredient(s) of a communicative event […] with which linguistic choices are interadaptable" (Verschueren 2008: 18). Besides the language users, the co-text, the inter-text and the communication medium, any context of language use comprises "[a]spects of physical, social and mental reality [that] get 'activated' by the utterer and the interpreter in their respective choice-making practices, and that is how they become part of language use as elements with which the making of choices is interadaptable" (Verschueren 1999: 87–88). In our methodology, the basic unit of analysis is not an isolated word or phrase, but a newspaper article. Moreover, our results are interpreted by taking more (social, political, historical, institutional) context into account than the co-text provided in concordance lines.

Nevertheless, it must be acknowledged that corpus linguistic studies with their focus on frequency and typicality (Stubbs 2001: 151) can provide interesting background information for discourse studies. In this respect, Krishnamurty's corpus-based study of the ideological meanings of the words *ethnic, racial* and *tribal*, is

particularly relevant for our research. He clearly laid bare the pejorative connotations and typical uses in negative contexts in the English language, particularly in newspapers (Krishnamurty 1996). Here it must be stressed that our promotion of text mining tools to study discourse does not preclude the supplementary use of tools which are associated with corpus linguistics, proof of which is our own use of *WordSmith* (see Section 4.2). Also interesting from our point of view are explorations in corpus linguistics of pragmatic phenomena, such as presuppositions, turn-taking, deixis or speech acts (e.g. Rühleman 2010; O'Keeffe et al. 2011). This could be a domain of synergy in the future.

2.2. *Related text mining approaches*

Text mining, a research field that has its roots in *data mining* (Witten and Frank 2005) and *machine learning* (Mitchell 1997), is a well-established technology for document analysis (Feldman and Sanger 2007). In particular, text classification methods are already routinely used for different types of newspaper article classification. The majority of tasks addressed are related to topic, genre and author classification (e.g. Cohen and Singer 1999; Liu and Hu 2007; Finn and Kushmerick 2006; Zhao and Zobel 2005; Stamatatos et al. 2000). An impressive application of statistical and machine learning approaches used in daily monitoring of news from different media is the Europe Media Monitor[1], a research and development effort of the European Joint Research Center in Ispra, Italy, that gathers reports from news portals world-wide in 43 languages, classifies the articles, analyzes the news texts by extracting information from them, aggregates the information, issues alerts, and produces visual representations of the information found.

Text mining approaches have already been introduced into ideology and opinion analysis. Balahur and Steinberger (2009), for instance, explore sentiment analysis in newspaper texts, aiming at discovering the positive or negative opinions expressed in the articles on a given topic. In case of newspapers, they argue, three different components are to be distinguished: The author, the reader and the text itself. Concentrating on analyses of quoted text, they established guidelines for positive and negative sentiment annotation. Fortuna et al. (2009) present an application of statistical learning algorithms to the analysis of patterns in media. They analyze two types of biases in four international online media: CNN, the English version of *Al Jazeera*, *International Herald Tribune* and *Detroit News*, in the period of March 2005 to April 2006. Their analytical focus was to find the bias in the choice of topics different sources report on, and to observe different choices of terms when reporting on a given topic. Lin et al. (2008) proposed a statistical model for ideological discourse, based on the hypothesis that ideological perspectives can be detected through lexical variations. On the Bitterlemons corpus[2], which aims at contributing to mutual understanding between Palestinians and Israelis through the open exchange of ideas, they observed that some words in discourse are used more frequently because of their relation to the text topic, while other words were used more frequently because of the author's particular ideological perspective. They encoded the lexical variations in ideological discourse in topical and ideological weights of words. In contrast to

[1] http://emm.newsbrief.eu/ [01/04/2011]
[2] More information can be found at http://www.bitterlemons.org/ [09/11/2010].

Fortuna et al. (2009), Lin et al. (2008) cover the topical and lexical/ideological aspect in a common model.

Our own approach differs from the text mining methods presented above in two respects. Firstly, none of the above approaches focuses on qualitative discourse analysis of the results. They provide lists of words or word types that are indicative for a certain type of ideologically biased discourse, e.g. racist/non-racist in Greevy and Smeaton (2004), Palestinian/Israeli in Lin et al. (2008), and American/Arab in Fortuna et al. (2009), but do not provide the interpretation of features from a discourse-oriented theory. Secondly, these approaches do not take into consideration different combinations of words.

3. Motivation: Studying aspects of ideology in news reporting

In this study we aim at capturing significant differences between local and Western press coverage of the 2007 Kenyan elections and the following post-election crisis using a combination of text mining techniques and pragmatic analysis. For our experiments, we selected newspaper articles from a larger corpus that was originally collected as part of the *Intertextuality and Flows of Information* project[3] in which an ethnographically-supported pragmatic analysis of news discourse is undertaken. The newspaper articles were culled from six English-language quality newspapers. For convenience sake – although running the risk of over-generalizing – the American and British newspapers *The New York Times*, *The Washington Post*, *The Times* and *The Independent*, will be labeled *Western*; this term does not denote a geographical nor geopolitical entity, but rather refers to a presumably similar ideological space in full awareness of its inherent heterogeneity and internal contradictions. In a similar simplifying move, in this paper *The Standard* and *Daily Nation* will generally be treated as *local*, thus ignoring several other Kenyan dailies.

In view of the discourse analysis carried out in this paper, we explain our expectations of finding ideological differences in national versus international news reports of the same events. Acknowledging that it is impossible to fully contextualize reality, which is always heterogeneous, multi-interpretable and complex, news can be regarded as a selective presentation of recent events that happen in the world and are deemed relevant or interesting for the audience. The adjectives 'relevant' and 'interesting' already refer to the evaluative aspect, typical of any news item. In general, news depends on institutional constraints, professional routines, conditions of information accessibility, specific journalistic and discursive choices, and a sense of relevance in relation to the idealized reader and news values (see also Section 7). Van Ginneken (2002: 36) defines news as something that is perceived as 'new' within a specific society or social group, something that is considered to be unexpected, extraordinary and abnormal. Crucially, what is deemed normal or irrelevant does not tend to be made explicit.

The common ground upon which the determination of news values is based or the worldview within which news discourse must be understood usually remains implicit. This makes news inherently 'ideological' in the sense that – like most types of

[3] This research project, carried out by Liesbeth Michiels and Roel Coesemans at the IPrA Research Center, University of Antwerp, studies ideological aspects of processes of (implicit) meaning generation and transformation in (inter)national newspaper reporting.

discourse – it carries along unquestioned assumptions. Newsmakers' interpretations and representations of newsworthy events are always made from a particular ideological position (e.g. Fairclough 1995; Ngonyani 2000; Richardson 2007). It follows that news is never a neutral representation of facts. As Reah (1998: 50) puts it: "Newspapers are not simply vehicles for delivering information. They present the reader with aspects of the news, and present it often in a way that intends to guide the ideological stance of the reader". Newspaper articles, as products of processes of meaning generation through choice-making, are also 'ideological' in that every choice made implies the rejection of other possible alternative choices (whether or not equally valid) that can lead to totally different meanings. In other words, "the linguistic choices that are made in texts can carry ideological meaning" (Fairclough 1995: 25). Matu and Lubbe (2007: 402) claim that "linguistic choices play a fundamental role in the propagation and perpetuation of implicit and dominant ideologies, and that there are certain ideological differences that are conveyed either tacitly or overtly in newspaper reporting".

It is beyond the focus of this paper to elaborate on the complex concept of ideology, but a minimal clarification is in order. Ideology will be broadly conceived of as "any constellation of fundamental or commonsensical, and often normative, beliefs and ideas related to some aspect(s) of (social) 'reality'" (Verschueren 1999: 238). This conception relates ideology to normality and worldview. It must be remarked that ideology consists of both implicit and explicit views on reality. Ideological content tends to be taken for granted when it relates to what is (thought to be) generally acceptable. When an interpretation of a newsworthy event is presented as inevitable and a news report is written on the basis of presumably natural assumptions, the underlying worldview will rarely be questioned, even though the interpretation and the news report as a whole might well be contestable from a different perspective. Note that our notion of ideology does not equal social cognition or socially shared belief systems in the sense of Van Dijk (2006), but we share with him the belief that ideology has social and cognitive functions, influencing how we think and act. To use Verschueren's words, discursively constructed ideological webs "serve the purpose of framing, validating or legitimating attitudes and actions" (1996: 592). From this perspective, we can interpret Harris' conviction that "[m]edia affect our minds [since they] give us ideas, change our attitudes, tell us what the world is like" and thus impact on the way we live (Harris 2004: 270).

Having established that newspaper articles are a matter of choice and ideology, and that they are both constituted by and constitutive of social reality, differences of reporting in various newspapers are to be expected, even if the same events are covered. For the purposes of this paper, we focus on the differences between the Kenyan and Western press coverage. The Kenyan and Western newspaper articles constitute different subgenres: the former fall into the category of national news, while the latter concern foreign or international newspaper reports. This means that a foreign correspondent has to adapt and write local news stories in such a way that they are easily understandable to the home audience. Often an angle of reporting is searched so that the foreign news can be rooted into the readers' background knowledge or anchored into a familiar frame of interpretation. That is why a newspaper report typically takes one (or at best a few limited) perspective(s) to an event, while other aspects of reality are obscured, underexposed or just missed. In this respect, Lee et al. (2000: 295) observed how "the same event may be given distinct media

representations by various nations through the prisms of their dominant ideologies as defined by power structures, cultural repertoires and politico-economic interests". They concluded that "[m]edia domesticate foreign news in the light of their own national interests and cultural assumptions" (Lee et al. 2000: 306).

In short, we expect fundamental differences between the Kenyan and the Western part of the corpus because the articles are written for different target audiences. Through our analysis, we will not get to the bottom of ideology or implicit meanings, rather our aim is to use automated text mining techniques in order to explore where lexical choices differ in the Western and local written media, and how the discovered patterns of lexical variations can imply ideological differences, thus supporting a further pragmatic analysis of the newspaper discourse in the Western and local media.

4. Introduction to the corpus

In our study we analyzed 464 articles, spanning a time period from December 22, 2007 to February 29, 2008. As the four selected Western newspapers (WE: *The New York Times*, *The Washington Post*, *The Times* and *The Independent*) published 232 articles on the topic of the Kenyan elections and crisis, we selected an equal number of 232 articles also from the local dailies (LO: *The Standard* and *Daily Nation*) in order to have a symmetrical corpus in terms of the numbers of articles of the two classes (WE and LO).[4] In this section we first present the background of the Kenyan election crisis (Section 4.1), followed by initial data exploration using a standard corpus linguistics tool (Section 4.2).

4.1. *Historical and political background on the Kenyan elections*

For a better understanding of the corpus, the news texts will be briefly situated in their historical and political context. The setting of the events is the Republic of Kenya, a multi-ethnic country with a booming economy and a centralized government where much of the power resides in the president. The incumbent, Mwai Kibaki, is only the third president since Kenya gained independence from Britain in 1963. Despite a growing urban middle class, the gap between the rich elite and the poor masses remains wide. Poverty, unemployment, occasional periods of drought and unequal distribution of power and natural resources over Kenya's people regularly causes tensions, especially in the city slums.

Expectations ran high when Kibaki won the elections in 2002 as the presidential candidate of the National Rainbow Coalition, thanks to the support of the businessman and influential opposition politician Raila Odinga. It was the first time since Kenya had become a multiparty state in 1991, when the main opposition parties joined forces

[4] While in corpus linguistics it is common practice to compile two corpora of similar length, measured by the number of words, the main technology used in this article is text mining where analysis units are articles and not individual words as in corpus linguistics. As an asymmetrical corpus could represent some difficulties for some data mining techniques (especially for decision tree algorithms used in Section 5.2), we decided to make the corpus symmetric in terms of the numbers of articles of each of the two classes (*Western* and *local*). The selection of local articles was performed such that the two local newspapers were equally represented for the whole period. Within these constraints a random selection was made (from the same date we randomly selected an article).

to remove from power the Kenya African National Union. Although Kibaki succeeded in boosting the economy and installing free primary education, he failed to provide equal access to vital resources and reneged on his promise to reduce the power of the presidency by creating the post of prime minister for Odinga. In 2005, the latter left the government out of disagreement with the failed constitutional reform process. Together with some other dissidents, he founded the Orange Democratic Movement (ODM). Through the subsequent reshuffle, Kibaki's government, which had already been weakened by major corruption scandals, now lost its ethnic diversity and came to be perceived as an organ of cronyism (cf. Ogola 2009).

Even though Kibaki had lost a lot of credit, the 76-year-old president stood up for re-election within the Party of National Unity (PNU). In the run-up to the elections politics became increasingly polarized. During the campaign aggressive rhetoric was not eschewed and the ethnic angle was ever present (Rambaud 2008). Partly this was a result of some problematic characteristics of Kenyan politics, partly it was due to specific campaign strategies. In Kenya, political parties are seldom based on ideology, rather on social cleavages, as numerous politicians "are not motivated by party principles or constructive policy commitments", but instead "are more concerned with the quest for raw power, perceived as attainable by relying on the ethnic card" (Oloo 2007: 111). Moreover, in the "single-member-district first-past-the-post winner-takes-all" electoral system ethnic support is crucial (Oloo 2007: 121). In this respect, it is useful to know that Kibaki is a member of the Kikuyu ethnic group, which makes up 17% of the population, while his main contender, Raila Odinga, belongs to the ethnic group of the Luo, representing 10% of all Kenyans. More reasons for the ethnicization of the 2007 elections can be found in the campaigning. Simply put, ODM presented itself as a coalition of minority tribes (though dominated by Luo, Kalenjin and Luhya) that stood up against "Kibaki's Kikuyu government". It promised an equal distribution of wealth by a tribally-mixed, corruption-free government in a reformed federal state. While Odinga rocketed in the opinion polls, tensions rose when Kibaki not only personally installed five new judges to the Court of Appeal, but also appointed 19 of the 22 commissioners of the Electoral Commission of Kenya (ECK), which was interpreted as "a means through which he would use state institutions to stay in power" (Ogola 2009: 61).

The General Election, comprising presidential, parliamentary and civic elections, took place on 27 December 2007. Up to 72% of the eligible voters went to vote. The swiftly processed civic and parliamentary results indicated that people had opted for change by voting for novices or underdogs irrespective of their party or ethnicity. The National Assembly became dominated by ODM with 99 of the 210 parliamentary seats, while PNU only obtained 43 seats. Also in the civic polls ODM triumphed.

By contrast the outcome of the presidential results took unusually long. Anxiety grew as concrete evidence of fraud reinforced widespread rumors of rigging and the ECK lost control of the tallying process. On Friday December 28, it looked like Odinga was winning with a lead of one million votes, but the difference with Kibaki narrowed overnight to 38,000 votes. The following day the tallying was cancelled due to protests and conflicts between party members and ECK-officials, after which observers and media were thrown out of the tally centre by the paramilitary police. Most disputes revolved around fraudulent augmentation of votes and unrealistic voter turnout.[5] With an incomplete tally and available results lacking the required statutory

[5] For example, for the constituencies of Molo (Rift Valley Province) and Kieni (Central Province),

documents, ECK boss Samuel Kivuitu released final results on Sunday 30 December 2007. Mwai Kibaki of PNU was declared the winner with 4,584,000 votes; Odinga of ODM was said to have 4,352,000 votes. According to the disputed results Kibaki won a majority of votes in four provinces (Central 96.4%, Eastern 49.8%, North Eastern 50.9%, Nairobi 47.3%), while Odinga received most votes in the other four provinces (Nyanza: 82%, Western: 66.5%, Rift Valley 64.1%, Coast 58.8%). Different observer groups, including the East African Community Observer Mission, the Kenya Elections Domestic Observation Forum and the Commonwealth Observer Group, branded the presidential elections as deeply flawed. The European Union Election Observer Mission to Kenya concluded that these elections "leave a legacy of uncertainty as to who was actually elected as President by the Kenyan people", resulting in "an unprecedented situation in the country characterized by deep ethnic rifts and civil unrest as well as a political stand-off" (EU EOM 2008: 37).

The outcome of the presidential elections immediately triggered mass demonstrations by opposition supporters, but also rioting by frustrated youths, looting by criminal gangs and excessive use of force by the police in response. When on New Year's Day the ECK chairman Kivuitu publicly admitted that he was not sure whether Kibaki had won the elections, popular anger grew and chaos spread. Most outrages took place in and around the slums of five provinces: Central, Nairobi, Nyanza, Rift Valley and Western. This hints at the importance of the socio-economic local context of the violence during the crisis. In general three main categories of violence could be distinguished: spontaneous violence as a result of the elections and the political deadlock, organized attacks against targeted communities following unresolved disputes or long-standing grievances (e.g. about land rights), and organized retaliations. The Kenya National Commission on Human Rights reported instances of political violence, violent protest, criminal acts of killing, looting and destruction of property, pre-planned ethnic violence, and sexual and gender-based violence. So the violence cannot be uniformly labeled. Without denying that in some regions political protest turned into ethnic violence or was abused to settle tribal scores, it is clear that the ethnic is only one aspect, which cannot be plainly generalized (see the interpretations in Section 7).

Eventually, it took a lot of national and international pressure and mediation to resolve the political stalemate and end the societal crisis. On 28 February 2008 chief mediator Kofi Annan brokered a power-sharing deal. A total of 40 ministers, equally taken from ODM and PNU, were sworn in on 17 April 2008, when president Mwai Kibaki's cabinet finally became operative with Raila Odinga as prime minister. However, the political climate remains volatile. There is still a lot of disagreement about the inevitable constitutional reform process, and tensions are ever present, not only between the partners of the coalition government, but also within the ruling parties and between the supporters of the different fractions. Up to 1,200 Kenyans died as a direct consequence of the post-election crisis and more than 300,000 people lost their homes.

Kibaki had 20,000 and 17,000 more votes, respectively, in the final announcement of the results at the ECK headquarters in Nairobi, compared to the results announced on the spot by the returning officers in the presence of EU observers (EU EOM 2008: 34).

4.2. *Initial insights into the corpus*

In order to get a first impression of the data, we first performed a simple keyword frequency analysis using the WordSmith lexical analysis software for finding patterns in text (Scott 2008). Table 1 presents the first 30 keywords of the local and the Western part of the corpus. The list is sorted by the *keyness* value, a measure that compares the relative frequencies of a word in the given text compared to a reference corpus: a word that is frequent in the given text and rare in the reference corpus gets a high keyness value. As the reference corpus, we used the British National Corpus (BNC)[6]. Given the space limitations we present only the keyness value, together with the absolute and the relative frequency in the given text. In addition, the table presents the measure of dispersion of individual keywords: the Julliand's D index that indicates how uniformly the word is distributed in the corpus: high values mean that the words are distributed evenly throughout the corpus.

The most interesting observation is that in the local press we do not find explicit references to ethnicity, while in the Western press the word that gets the fifth highest keyness score, *Kikuyu*, is an ethnic tag. Other ethnicity-related words frequently used in the Western press are: *Kikuyus* (N11), *ethnic* (N15), *Luo* (N21), *Kalenjin* (N23), *tribe* (N24), and *Luos* (N27). In contrast, the words that typify the local press belong to the lexical field of (Kenyan) politics. The acronyms of the major political parties, ODM and PNU (N1, N10), have a high keyness value. The same holds for ECK (Electoral Commission of Kenya) and MPs (Members of Parliament). Also striking is the high frequency of *mediation* (N15) and *talks* (N17), especially in comparison to their absence in the top keywords of the Western press. As we will see later, it is no coincidence that the 25th keyword is *political* in the local press, while it is *tribal* in the Western press.

Note, however, that frequency-based analysis can be sometimes misleading as a purely quantitative and highly de-contextualized representation of a (corpus of) text(s), and thus requires prudent interpretation. For instance, the frequency of ODM (N1) in the local part of Table 1 as opposed to the lower frequency of PNU (N10) might suggest that the party ODM was covered more in the Kenyan newspapers than the PNU. In reality the contrary was true. A classical content analysis revealed that the PNU received 54% share of coverage in the *Daily Nation* compared to 55% in *The Standard*, while ODM got 29% in the former newspaper and 30% in the latter (EU EOM 2008: 26). The explanation of this misrepresentation lies in the fact that the third largest party in the elections was called ODM-Kenya. This faction split off from the larger Orange Movement and later leaned towards PNU, so that its leader Kalonzo Musyoka could become Vice-President. As WordSmith does not count concepts, but word tokens, the ODM label refers to all instances of ODM, whether separately or in the composition ODM-K(enya).

[6] World corpus is available at http://www.lexically.net/downloads/BNC_wordlists/downloading%20BNC.htm [01/06/2011].

	Local (LO)					Western (WE)				
N	Keyword	Keyness	Freq.	%	Disp.	Keyword	Keyness	Freq.	%	Disp.
1	ODM	10252,93	795	0,51	0,92	KIBAKI	10478,96	827	0,50	0,90
2	KIBAKI	8295,84	651	0,42	0,89	KENYA	9500,88	943	0,57	0,94
3	KENYA	5732,61	602	0,38	0,93	ODINGA	9362,28	755	0,46	0,89
4	RAILA	5409,67	420	0,27	0,79	KENYA'S	6091,10	513	0,31	0,90
5	ANNAN	4696,75	402	0,26	0,79	KIKUYU	5542,06	451	0,27	0,83
6	MR	4256,45	1319	0,84	0,96	ELECTION	4300,16	757	0,46	0,91
7	ODINGA	3733,38	308	0,20	0,87	KENYAN	3672,08	341	0,21	0,95
8	PRESIDENT	3661,46	754	0,48	0,89	OPPOSITION	3298,01	608	0,37	0,87
9	KENYANS	3353,50	273	0,17	0,89	MR	3213,25	1128	0,68	0,80
10	PNU	3215,23	249	0,16	0,77	VIOLENCE	3190,29	527	0,32	0,89
11	VIOLENCE	3179,73	519	0,33	0,86	KIKUYUS	3125,44	244	0,15	0,83
12	ELECTION	2498,46	500	0,32	0,90	RAILA	3086,84	242	0,15	0,95
13	NAIROBI	2495,77	247	0,16	0,91	NAIROBI	3067,93	299	0,18	0,92
14	SAID	2399,71	1510	0,96	0,96	KIBAKI'S	2907,66	227	0,14	0,86
15	MEDIATION	2264,43	232	0,15	0,84	ETHNIC	2794,59	393	0,24	0,78
16	PRESIDENTIAL	2075,46	301	0,19	0,91	KENYANS	2644,00	219	0,13	0,87
17	TALKS	2041,88	390	0,25	0,77	MWAI	2438,22	194	0,12	0,91
18	CRISIS	1844,72	350	0,22	0,84	PRESIDENT	2404,46	569	0,35	0,92
19	ECK	1701,01	146	0,09	0,67	ANNAN	2377,22	216	0,13	0,82
20	KOFI	1674,52	134	0,09	0,83	SAID	2168,71	1475	0,89	0,93
21	GOVERNMENT	1567,30	676	0,43	0,87	LUO	2125,23	176	0,11	0,80
22	LEADERS	1494,84	318	0,20	0,90	ODINGA'S	2042,20	164	0,10	0,77
23	KENYAN	1345,02	139	0,09	0,88	KALENJIN	1580,49	125	0,08	0,80
24	KIVUITU	1291,16	100	0,06	0,55	TRIBE	1543,66	193	0,12	0,83
25	POLITICAL	1273,01	463	0,30	0,89	TRIBAL	1431,85	182	0,11	0,89
26	YESTERDAY	1271,49	383	0,24	0,92	RIFT	1393,17	164	0,10	0,77
27	POLICE	1194,51	427	0,27	0,80	LUOS	1320,77	104	0,06	0,75
28	KISUMU	1157,63	94	0,06	0,68	KOFI	1229,14	100	0,06	0,80
29	MPS	1119,01	195	0,12	0,82	LEADERS	1146,89	268	0,16	0,80
30	ELDORET	1116,15	89	0,06	0,88	VOTE	1071,38	254	0,15	0,82

Table 1: Keywords obtained by the WordSmith lexical analysis software.

5. Detecting contrast patterns through classification model construction

In this section we use text mining methods to learn interpretable classification models which will allow us to study the differences in local and Western news reporting. This section is divided into three subsections. Section 5.1 is devoted to data representation and presents the transformation of documents into a feature vector format, Section 5.2 outlines document preprocessing and the experimental setting, also explaining the used text mining methods through a simplified example, and Section 5.3 presents the results of our experiments.

In text mining, document classification (or categorization)[7] refers to the task of classifying a given document into one or more categories based on its contents (Sebastiani 2002). To enable automated document classification, a classification model (*a classifier*) needs to be learned from the data. Text mining methods are either *supervised* or *unsupervised*. *Supervised learning* (which is the topic of this section) is performed as follows. Given a set of documents, pre-classified into distinct classes or categories, the goal of text mining is to automatically construct a classification model that will enable to assign one of the predefined categories to a new text document. Several text mining tools enable the construction of understandable classification models (for instance, in the form of a set of classification rules) that describe the categories and their differences. On the other hand, *unsupervised learning* (addressed in Section 6) is performed entirely without reference to external information, e.g. by document clustering and in this way determining distinct document categories.

In the analysis of articles on Kenyan elections, performed in this section, our actual goal is not to classify articles into one of the two classes, *local* and *Western*, as the categories of articles are clearly defined by the source from where the articles were taken. Instead, from the given class-labeled newspaper articles we automatically construct classification models with the goal to describe the two categories and to improve the understanding of their differences.

5.1. *Data representation*

In text mining, one of the key issues is the representation of text documents in such a format that captures their meanings in a compact way, and at the same time enables efficient processing by text mining algorithms. Such data preprocessing can be described as the transformation of unstructured text documents into a structured computer-readable representations. The main task of data preprocessing is to identify and extract representative *features*, e.g. words or combinations of words, and to represent each document by the features which characterize the document, e.g. the set of words which appear in the document (Feldman and Sanger 2007).

In data preprocessing, documents are transformed into the so-called *feature vector* format, where each feature vector corresponds to one newspaper article. The most frequent representation of feature vectors is the *bag of words* representation, where individual features correspond to individual words. In this approach, a document is represented as a vector of features with as many components as there are different words in the corpus. Every component represents a word in its normalized form, and if this word occurs in the document, the value for this feature vector component is set to 1, otherwise the value is set to 0.[8]

To explain the *bag of words* feature vector representation, take an illustrative example of two simplified documents consisting just of the article titles:

[7] In this paper we use classification (used in data mining and machine learning) and categorization (used in text categorization) as synonyms.

[8] Values 1 and 0 are known as binary feature values. Alternatively, term frequency or other similar values, such as TF-IDF (term frequency inverse document frequency) can also be used. We will use values 1 and ‘yes’ interchangeably to denote the presence of a word in the document, and 0 and ‘no’ for its absence.

(1) *Kenyan elections in chaos*
(2) *ECK delays results to avoid chaos*

In the standard procedure of feature vector construction, less informative words like *in* and *to* (the so-called stopwords) are removed, and the remaining words are normalized (lemmatized). In our example, the list of remaining normalized words is the following: [*avoid, chaos, delay, ECK, election, Kenyan, result*].

(1) The first article, which contains the words *chaos, election, Kenyan*, is represented by feature vector [0, 1, 0, 0, 1, 1, 0].
(2) The second article, containing the words *avoid, chaos, delay, ECK, result*, is represented by feature vector [1, 1, 1, 1, 0, 0, 1].

Note that the word *chaos* occurs in both documents, represented by 1 at the second place in the bag of words vectors.

Instead of individual words (word *unigrams*, W1), document representation can also be based on combinations of words such as word *bigrams* (sequences of two words, W2) and *trigrams* (sequences of three words, W3).

5.2. *Data preprocessing and experimental setting*

Since our aim is to better understand the way of reporting on the same event by the local and Western media, all the information that could be distinctive for the two classes but is irrelevant for our analysis was removed from the articles. To illustrate this point, newspapers normally have only few journalists covering Kenyan affairs, so if bylines were not removed, the author's name could easily be selected as a distinguishing feature. We therefore first performed data cleaning by removing meta-information such as newspaper source, authors of articles, dates of publication, photographers, mail addresses of authors, types of articles, etc. and used only the remaining relevant data[9] (headlines, crossheads, text of the article and photo captions).

As our experiments are aimed at discourse analysis, we did not want to exclude any type of words beforehand in data preprocessing, since we consider all word types potentially important. Therefore, we opted for experimenting with the data without performing stopword removal, lemmatization or stemming, techniques often used in document preprocessing. In our corpus of 464 newspaper articles, which comprises about 320,000 words, we considered word unigrams (W1) as well as word bigrams (W2) and trigrams (W3) as features in our experiments. We used binary valued features, calculated on the basis of the presence (value 1) or absence (value 0) of the term in an article. However, since using all the features would result in too large feature vectors, automated feature selection of the best 500 features[10] was done using

[9] Keeping the headlines, crossheads, pull quotes and captions means that with these news discourse specific phenomena we often included repetitions of words when counting word frequencies. As we believe that these repeated words also grasp readers' attention we have decided not to exclude them. Anyhow, in the experiments described in this section, word frequencies do not play a role, as the binary representation of documents was used.

[10] The significance of words has been ensured in data preprocessing by using the chi-2 test (only the 500 most significant words were used for document representation).

the TACTiCS system (Luyckx, 2010).

Three data mining algorithms were used for classification model construction: two rule learning algorithms (JRip and PRISM) and one decision tree learning algorithm (J48). The choice of these symbolic data mining algorithms from the Weka data mining toolkit[11] (Witten and Frank 2005) enabled us to extract interpretable classification models (in the form of sets of rules and decision trees) from the document corpus. The choice of rule learning and decision tree learning algorithms was motivated by the need to ensure simple interpretability of the results.

To explain the idea of how classification models are learned from text documents and how to use the models for explaining the differences between two text corpora, take the following simplified example. Suppose we have 100 articles: 50 articles describing movie reviews, and 50 articles describing classical music reviews. In text preprocessing, the documents are transformed into a feature vector representation, where features correspond to individual words occurring in both sets of documents. A rule learning algorithm typically results in a set of rules for each class, with a class label (*movie* or *music*) in each rule conclusion, and a combination of most distinguishing or most characteristic features in rule conditions. An illustrative rule for the movie document category could be:

IF *actor=1* AND *screenplay=1* AND *symphony=0* THEN *Class=movie.*

This rule can be interpreted as "If the article contains the words *actor* and *screenplay*, and does not contain the word *symphony*, the article is from the *movie* category of articles". Take another example for the classical music reviews document category.

IF *orchestra=1* THEN *Class=music*.

Inspecting these two rules enables you to explore the differences between the two classes of documents. As opposed to these two artificially constructed rules, see the real examples of rules learned from the articles on Kenyan elections in Tables 3 and 4, and decision trees in Figures 2 and 3.

In general, rules are constructed in the following way. First, for a given class, the 'best' feature is selected based on a statistical test, followed by conjunctively adding other features, until the statistical stopping criterion for rule construction is satisfied (the exact description of the tests is beyond the scope of this paper). In the hypothetical movie and classical music reviews example above, the algorithm would select a class (*movie*), add a first condition to the rule (*actor=1*) and continue conjunctively adding features (words) until a significant group of documents is covered by the constructed description which separates this group of documents from the documents of the other class (*music*).

Two rule learning algorithms were used in our experiments: JRip and PRISM. The *JRip decision rule learning algorithm*, which implements the original algorithm of Cohen (1995), proceeds by selecting the examples of a given class and finding a set of rules that cover all the instances of this class. Each individual rule is automatically constructed as a conjunction of features (words, their bigrams or trigrams) which best characterize the given class. Thereafter it proceeds to the next class and does the same. The *PRISM rule learning algorithm* (Cendrowska 1987) generates only correct or 'perfect' rules for each class: it measures the success of a rule by the accuracy and any rule with accuracy less than 100% is considered 'incorrect'.

In addition, the *J48 decision tree learning algorithm*, which implements a decision tree learner developed by Quinlan (1993), was used. The algorithm builds a decision

[11] *Weka 3.6.0:* http://www.cs.waikato.ac.nz/ml/weka/ [09/11/2010].

tree consisting of nodes that correspond to individual features (words, word bigrams or trigrams), and arcs corresponding to tests (e.g. does the selected word occur or not). For illustration, see a simple, manually drawn decision tree for the movie and classical music reviews example in Figure 1.

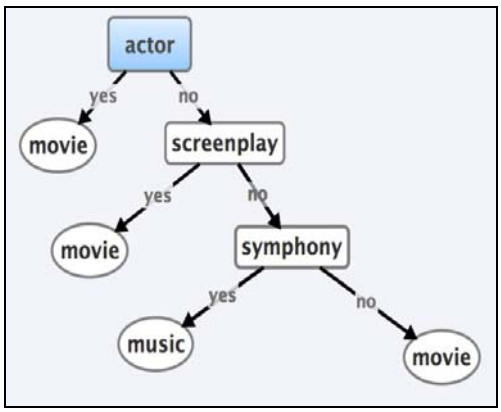

Figure 1. A simple decision tree.

The construction of a decision tree is a recursive process. First, the most informative feature is selected as the root node. Then, branches for each possible value of this feature are made (for our problem the tree will be binary, each feature represents a word or word combination that can be present or not). This process is repeated for each branch. The learning algorithm decides which feature to include as a node of the tree by calculating the informative value of a feature by an information-theoretic method called *information gain*. When using the tree for classification, depending on the value of a feature in the given node, a different sub-tree is accessed, or a different output class is assigned to the instance (document) being classified.

5.3. *Experimental results and analysis of classification models*

Table 2 presents the quality of the constructed classification models in terms of their classification accuracy.[12] The evaluation of classification accuracy (i.e. accuracy of classifying an article as being of class LO or WE) was done with 10-fold cross validation[13], which means that we did each experiment ten times with different training and test sets, reporting the averages and standard deviation. Given that there is the same number of articles in each of the two classes, the baseline accuracy of predicting randomly the origin of a news article is 50%. The results show that all the learning algorithms perform very well, e.g. the J48 and JRip learning algorithms reach very high classification accuracy of about 90%.

[12] Reporting on the accuracy of classification can be understood as means for estimating the reliability of automatically constructed models used as a starting point for our analysis. Although high classification accuracy of the constructed models is not our ultimate goal, high accuracy is desired as more accurate models also better describe the domain.

[13] 10-fold cross-validation is a standard evaluation procedure used in data mining and text mining. In each run of the evaluation, a training set of 90% of articles is used for learning the classification model, and the remaining 10% are used for its evaluation.

	Classification accuracy in % (and standard deviation)		
Feature set	J48	Jrip	PRISM
W1 (unigrams)	89.00 (4.64)	89.22 (3.95)	83.61 (6.34)
W2 (bigrams)	89.46 (4.17)	90.53 (3.81)	82.53 (4.41)
W3 (trigrams)	89.67 (6.03)	90.96 (6.34)	83.86 (5.58)

Table 2: Results of 10-fold cross-validation.

We now interpret some of the most interesting models and show that if we look for the context of the selected features in the corpus, we can get interesting additional information. We first present two different classification models automatically constructed with JRip. The first one (Table 3) was built in order to find rules for category *Western*, and the second one (Table 4) for category *Local*. In both cases, an ordered set of rules is constructed. The rules are constructed in such a way that the documents, which are covered by the currently constructed rule, are removed from the dataset before the next rule is constructed. The rules are ordered according to their *support* (the number of covered examples that fulfill the conditions of the rule). As the first rules are more representative than the others, they are more interesting for the interpretation.

	Covered by the rule	Correctly classified documents	Incorrectly classified documents
1. If the article contains the words *mwai* and *opposition*, and not the word *odm*, the class is Western	142	140	2
2. If the article contains the word *kikuyu*, the class is Western	48	45	3
3. If the article contains the word *raila*, but not the words *odm* and *about*, the class is Western	12	12	0
4. If the article does not contain the words *mr*, *will*, and *at*, the class is Western	9	9	0
5. If the article contains the word *club*, the class is Western	7	5	2
6. If the article contains the words *opposition* and *least*, the class is Western	9	7	2
7. Otherwise, the class is Local	237	223	14

Table 3. JRip results for class WE, constructed from 464 documents (W1 feature set); accuracy: 89.2%. The rules were transcribed from the original JRip format into natural language.[14]

The first rule, which holds true for 140 documents of class Western, states that if Kibaki's first name *Mwai* is used in an article where also the word *opposition* is present, while there is no reference to the abbreviated name of the main opposition party ODM, then the article originates from one of the Western newspapers. Also later classification rules suggest that in the Western press the term opposition is preferred

[14] For completeness, the first rule of Table 3 in the original JRip output format is provided: (feature43#mwai = 1) and (feature1#odm = 0) and (feature4#opposition = 1) => class=WE (142.0/2.0).

above the party name ODM (e.g. rules 3 and 6). If the first rule does not hold for an article to be classified, the rather simple second rule comes into play. This rule means that the use of the word *Kikuyu*, the name of Kibaki's ethnic group, is an indicator of the Western class. From this rule the tentative hypothesis arises that references to ethnicity are a distinguishing feature of the Western press. It could be argued that the Western press puts the news events into a tribal frame, because, from an ideological point of view, this was an easy or habitual frame of interpretation for both foreign correspondents and their readers in the US and the UK.

Let us look at other classification models to check whether this hypothesis holds and whether we can find other typical features of Western or local newspaper coverage about the Kenyan elections and the post-election crisis. Our observations and hypotheses following from these models will be further interpreted and put into perspective in Section 7 and in the conclusions of this paper.

	Covered by the Rule	Correctly Classified Documents-	Incorrectly Classified Documents-
1. If the article contains the words *odm* and not the words *opposition* and *corruption,* the class is Local	124	124	0
2. If the article contains the word *mr* and not the words *mwai* and *tribal*, the class is Local	77	69	8
3. If the article does not contain the words *odinga* and *Kenya*, the class is Local	13	11	2
4. If the article contains the words *dispute* and *odm*, the class is Local	6	6	0
5. If the article contains the word *crisis* and not the words *odinga* and *kikuyu*, the class is Local	13	9	4
6. Otherwise, the class is Western	231	218	13

Table 4. JRip results for class LO, constructed from 464 documents (W1 feature set); accuracy: 90.3%.

The first rule in Table 4 indicates that articles containing the party name ODM but missing the word *opposition* as well as the word *corruption* belong to the local media. Compared to the Western media (see the rules in Table 3), the Kenyan newspapers use the specific name of ODM, rather than the more general term *opposition*. The next most important rule dictates that if there is a feature *Mr* but not the features *Mwai* nor *tribal*, the article can be traced back to the local media. Here the absence of the feature *tribal* in the combinatorial rule for the local press is conspicuous. Moreover, both classification models (Table 3 and Table 4) indicate that references to concrete tribes, such as the Kikuyu, are typically present in the Western but absent from the local newspaper articles. This is an interesting observation as words like *tribe* and *tribal* are not ideologically neutral terms, but are frequently used for stereotyping diverse African societies and their conflicts (Ray 2008). As said before, we will come back to these issues in Sections 7 and 8 where more substantial interpretations will be provided. The third rule may seem curious, but it has a simple explanation. In many Kenyan newspaper articles the word *Odinga* is wanting, because in Kenya he is known by his first name. Frequently the proper name *Raila* is used to avoid confusion with his father, Jaramogi Oginga Odinga, who was a prominent politician, and with his brother who is also into politics. Note also that according to this classification model, the

conflict in Kenya is generally conceptualized by the local press as a dispute or a crisis.

The next experiment was done using the J48 decision tree learning algorithm, resulting in the classification model presented in Figure 2. In decision tree modeling we do not get a separate model for the Western and the local class. In the top node of the decision tree we find the feature *Kikuyu*. The left branch of the decision tree indicates that this word is frequent in the Western media, but that *Kikuyu* rarely occurs in the Kenyan part of the corpus (just in three articles). This classification model confirms our hypothesis: the Western press frequently refers to the ethnicity of the participants of the news stories, while this information is missing from the local newspaper articles. For further proof, we also checked the distribution of *Luo*, Odinga's ethnic group, and found that its distribution is very unbalanced as well (80 Western and 3 local press articles).

The further interpretation of the decision tree shows that if the word *Kikuyu* is not present, the next node gives feature *ODM* as an indicator of the local class. We also verified in the corpus that the full name of the party, *Orange Democratic Movement*, is more present in the local than in the Western media. Between other alternative lexical choices of naming ODM, we discovered twelve references to ODM as *Odinga's party* while in local coverage this expression appears only once. We can conjecture that the Western press reported the elections and the post-election crisis as a battle between tribes, assuming a sharp Luo-Kikuyu distinction, while the local press framed the events more in sociopolitical terms, explicitly downplaying or avoiding ethnic oppositions.

Finally, if *Kikuyu* does not appear and nor does *ODM*, but the first name of the ODM presidential candidate is part of the article, then it concerns a Western newspaper article. This does not contradict our earlier explanation about the frequent use of Odinga's first name in the local press. It rather reveals that the Western media (when not concentrating on tribal repartitions) almost exclusively focused on the main candidates of the presidential elections, while the Kenyan part of the corpus contains a lot of articles about other people, such as other politicians, community leaders, election commissioners or other officials, civil society spokespersons, etc. However, this information cannot be read off from Figure 2. Rather it results from pragmatic, contextual knowledge.

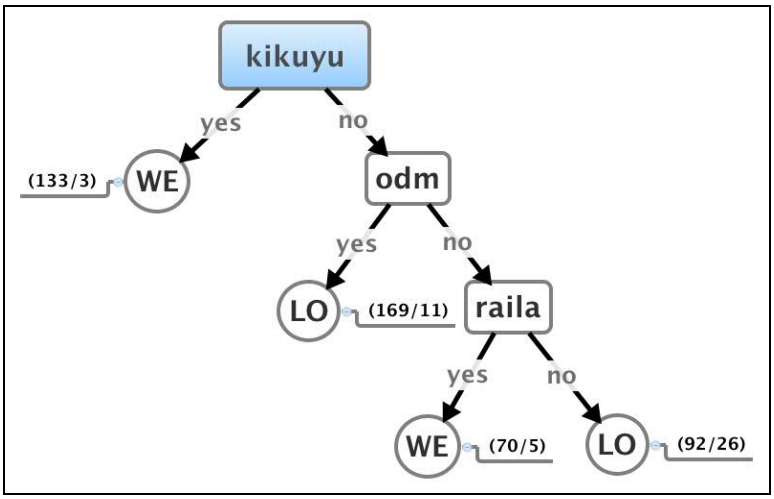

Figure 2. J48 results (464 documents, W1 feature set), accuracy 89%
(Number of articles covered by the rule / Number of incorrectly covered articles).

The decision tree of Figure 3, constructed on word trigrams, correctly classifies

nearly 90% of the examples, where the majority of the generated features concern the representation of the main protagonists. The referring expression *Mr Raila Odinga* proves to be the usual representation of the principal opposition leader in the local media. When Odinga is not referred to as *Mr Raila Odinga*, the decision tree generated a similar trigram, *Odm leader Raila*, as the second most important feature to distinguish between the local and Western newspaper texts. For articles lacking the latter trigram, the next distinguishing feature is *President Mwai Kibaki*. This third node in the decision tree model visualizes that an article is of the Western class when there is no reference to the opposition leader as *Mr Raila Odinga*, nor as *ODM leader Raila*, and when the elected candidate is presented in the news discourse as *President Mwai Kibaki*. In case Kibaki is not addressed in this way, but his full name is used, followed by a comma (*Mwai Kibaki*,), this is again an indication of Western journalistic writing. Descending to the final node, the model indicates that newspaper articles with *Raila Odinga* followed by a comma (*Raila Odinga*,) derive from the Western media too. So this particular decision tree facilitates the study of the representation of the social actors in a large corpus of newspaper articles.

These observations are significant in the sense that they suggest that in the local press Odinga is treated as one among many politicians, although an important and influential one. He is mainly neutrally referred to as *Mr Raila Odinga*, with an honorific introducing his full name comparable to, for instance, Mr Kalonzo Musyoka (presidential candidate for ODM-K) or Ms Nazlin Omar (presidential candidate of the Workers Congress Party). Or else he tends to be portrayed in the local press as one of the leaders of ODM or as its presidential candidate. His party is specified, rather than generalized as 'the opposition' (see also Tables 3 and 4 above). This is in marked contrast with the Western press, where Odinga is usually represented as the figurehead of the opposition, rather than as the presidential candidate of one of many opposition parties. Such a generalization could give the reader a distorted picture of Kenyan politics. After all, apart from Kibaki and Odinga, there were seven more presidential candidates and in total 159 parties participated in the General Election. Also of importance is what follows the comma in the last two nodes. As this cannot be read off from the model, we will return to it in Section 7.

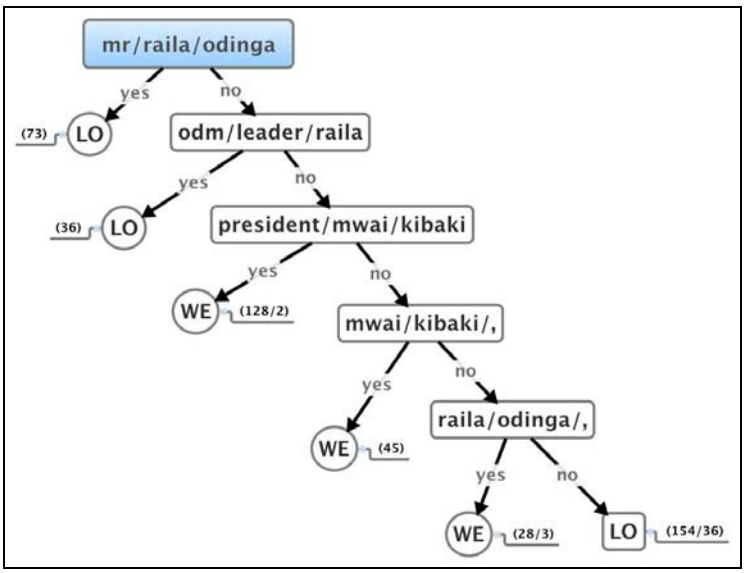

Figure 3. Results of J48 on W3. Accuracy: 89.6%.

Another model, given in Table 5, was generated with the PRISM algorithm, using

the trigram feature vector representation. Again we can notice quite a lot of the classifying selections involving differences in the use of ethnicity-related terms (either using the words *tribe, tribal* or in relation to the name of the *Kikuyu tribe*, as it can be seen from rules 2, 3, 6 or 17 for the Western class).

Apart from revealing the main tendency that the local press prefers a political perspective on the events while the Western press opts for a tribal frame (compare the many references to politics versus tribe), two additional observations are to be highlighted. First we want to point to the salient presence of Samuel Kivuitu, the chairman of the Electoral Commission of Kenya (ECK), in the local news discourse (see rules 17 and 7). Kivuitu features prominently in the Kenyans newspaper articles, as he is in charge of the organization of the elections, the tallying process and the announcement of the winner. In spite of his importance for the legitimization of the election results and his (and ECK's) role in the election crisis he is remarkably absent in the Western press. By often silencing his voice the Western media miss again a piece of the puzzle to better understand the Kenyan crisis. Secondly, it must be noted that not only the elections and the crisis but also the violence is depicted in tribal terms in the Western press. When the ethnicity of the social actors involved in the violence is not explicitly mentioned, Western journalists stress the primitive barbarity of what they see as tribal violence by their focus on exotic weapons, such as *machetes* (see rule 16 for the Western class) or *pangas* and *spears*, as revealed by additional analysis of the articles. Conversely, the Kenyan dailies speak of the *post-election* crisis and of *post-election* violence (rule 6). Instead of describing it as primitive and tribally-driven, they describe the violence in general terms as the *destruction of property* (rule 18 for the local class).

1.If the article contains 'mr/raila/odinga' then LO	1. If the article contains 'of/the/most' then WE
2.If the article contains 'odm/leader/raila' then LO	2. If the article contains ',/the/kikuyu' then WE
3.If the article contains 'a/press/conference' then LO	3. If the article contains "kibaki/'s/kikuyu" then WE
4.If the article contains ',/yesterday/.' then LO	4. If the article contains 'top/opposition/leader' then WE
5.If the article contains 'yesterday/,/the' then LO	5. If the article contains opposition/orange/democratic' then WE
6.If the article contains ' / /post-election' then LO	6. If the article contains 'the/kikuyu/tribe' then WE
7.If the article contains 'chairman/,/mr' then LO	7. If the article contains ' / /kenyan' then WE
8.If the article contains 'the/election/of' then LO	8. If the article contains "africa/'s/most" then WE
9.If the article contains 'affected/by/the' then LO	9. If the article contains 'a/kikuyu/.' then WE
10.If the article contains 'general/election/.' then LO	10. If the article contains 'the/dec./27' then WE
11.If the article contains 'he/,/however' then LO	11. If the article contains 'and/opposition/leader' then WE
12.If the article contains 'pnu/and/odm' then LO	12. If the article contains "'s/orange/democratic" then WE
13.If the article contains 'odm/and/pnu' then LO	13. If the article contains "kenya/'s/president" then WE
14.If the article contains ',/mr/john' then LO	14. If the article contains 'since/the/election' then WE
15.If the article contains 'to/the/media' then LO	15. If the article contains ',/mr/kibaki' then WE
16.If the article contains ',/mr/william' then LO	16. If the article contains 'armed/with/machetes' then WE
17.If the article contains 'mr/samuel/kivuitu' then LO	17. If the article contains "'s/kikuyu/tribe" then WE
18.If the article contains 'destruction/of/property' then LO	18. If the article contains 'killed/more/than' then WE
19.If the article contains './the/minister' then LO	

Table 5. Results of PRISM on W3 (only first few rules for each class), accuracy: 83.83%.

Running ahead of the contextual pragmatic analysis and discussion in Sections 7 and 8, it can already be acknowledged that phrases such as "divided on tribal lines, rival groups have been fighting with machetes and sticks" (from article '50 die in blazing church a specter of tribal war looms', *The Times*, 2 January 2008) or "more than 100 local Kalenjin militiamen armed with machetes and bows and arrows" (from 'Kenyans say tribal divide has reached police force', *Washington Post*, 12 February 2008), reveal an arguably ideological choice to present the post-election conflicts as primitive, tribal warfare, which has been criticized as an ideologically dangerous and colonialist perspective (e.g. Ray 2008). On the other hand, exclusions of certain actors, e.g. by means of nominalization as in the *destruction of property*, are ideological too, as is the backgrounding of relevant characteristics like ethnicity in those specific conflicts that took an ethnic dimension (cf. Van Leeuwen 2008: 28-32). This ideological choice is criticized for instance by Wrong who remarks that "the Kenyan media have essentially refused to cover the biggest story on their patch" by repudiating the notion of tribe, while she adds that "[y]ou cannot defuse a problem you refuse to see" (Wrong 2008: 23). We will come back to this issue in Section 8.

6. Contrasting keyword detection by semi-automated topic ontology construction

In computer science, the term *ontology* denotes a formal representation of a set of concepts of a domain and the relationships among these concepts. Ontologies are organized hierarchically: a concept is divided into a set of sub-concepts. A concept representing a set of documents can also be described by the main topics addressed in the documents. Accordingly, a *topic ontology* (Fortuna et al. 2007) is a hierarchical organization of documents' topics and their sub-topics. We used a semi-automated topic ontology construction tool OntoGen[15] (ibid.) mainly aimed at building topic ontologies by unsupervised learning from unlabeled data, but can be applied also for other purposes, such as classification of documents, document search, etc.

This section describes how a topic ontology was built from our set of articles and how the topic ontology construction tool was used to search for differences between the local and the Western press coverage of Kenyan elections. As input for OntoGen, we used the lemmatized document representation, where lemmas were obtained by using a memory based shallow parser (Daelemans et al. 1999).

In OntoGen, hierarchical decomposition of a given set of documents into document subsets is performed by *k*-means clustering[16], where *k* is defined by the user at each step of the multi-layer hierarchical ontology construction process, and each sub-domain (sub-concepts) is described by the main topics that the documents cover. OntoGen offers two different ways of getting the topic descriptions. One option is to get a list of keywords composed of the most contrasting words, where these distinguishing keywords are extracted by the Support Vector Machine (SVM) classifier.[17] The other option is to get a list of keywords composed of the most descriptive words for the document cluster: i.e. *n* most frequent keywords describing

[15] For more information, http://ontogen.ijs.si/ [14/09/2009].

[16] Clustering is an unsupervised learning method which groups documents into document clusters according to their similarity, where *k*-means clustering results in *k* different clusters.

[17] Support Vector Machine denotes supervised learning methods that build a classifier by constructing a separating hyper-plane in a high-dimensional space of features which has the largest distance to the nearest data points (documents) of the different classes (LOCAL and WESTERN).

the document cluster.[18]

In our experiments the first step of topic ontology construction was the same: the first step of document grouping was performed by manually enforcing the root node (representing all the documents) to be split into two clusters WESTERN and LOCAL (corresponding to Western and local newspaper articles, respectively). The root node in the center of the topic ontology (see Figure 4), which stands for the whole dataset, was thus split up into a Western and local subset, each containing 232 documents. By manually enforcing this separation into two document sets enabled us to compare distinct topics separately for each of the two classes.

In the first experiment we explored the contrasting view of the Western and local media coverage via the analysis of SVM-keywords.[19] After creating the local and Western class and categorizing the documents into these two categories, we analyzed the distinguishing keywords, uncovered by the SVM algorithm implemented in OntoGen. The SVM-based contrasting keywords, best distinguishing between the articles of the two classes, are presented in Table 6.

LOCAL	odm, mp, team, mr, pnu, odm_leader, president_kibaki, dr, media, statement
WESTERN	kikuyu, mr_kibaki, opposition, mr_odinga, luo, tribe, tribalism, opposition leader, odinga, ethnic

Table 6. SVM (contrasting) keywords for the local vs. Western articles.

These words show the differences in the way of referring to the main protagonists: in local articles the definite descriptions *ODM leader* and *President Kibaki* are used, while the main participants are described more often as *Mr Kibaki*, *Mr Odinga*, *opposition leader* and *Kikuyu* or *Luo* in the Western newspapers. Local media limit their coverage to political parties and functions: *ODM, PNU* and *MP*, while Western media present the election through an ethnic lens by making use of such words as *Kikuyu, Luo, ethnic*, and even the more ideologically marked words *tribe* or *tribalism*.

In the second experiment, the initial document groups, corresponding to the local and Western articles, were further automatically split into subgroups by using the OntoGen's *k*-means clustering algorithm (the parameter of k-means clustering was manually set to k=3), in this way forming lower-level concepts and sub-concepts (i.e. topics and sub-topics). The resulting topic ontology is shown in Figure 4.

OntoGen first identified three central topic domains which were further divided in three topic subdomains, each of which is described by a set of keywords representing the main topics of the article group. In that way the topic ontology represents the relations between the articles while making a link to their content. So from Figure 4 we can infer what is the articles coverage, and inspect the similarities between the related articles within each cluster and in contrast with the other document clusters.

Broadly speaking, the central topics distinguished for the local subcorpus clearly correspond to three key stages in our case study: the elections (*odm, eck, court*), violent protests or rioting (*police, media, mp*) and the search for a solution (*odm, Annan, talk*). The whole election process is extensively covered, from the organization by the ECK over the announcement of the results by ECK chairman Kivuitu to different problems during and immediately after the elections. The former problems led to critical reports from observers, whereas the latter concerning the outcome of the

[18] In more detail, these are the most important words describing the centroid (the artificial 'average' document) of the document cluster.

[19] For more details of the use of SVM for keyword extraction in OntoGen (see Fortuna et al. 2006).

elections prompted coverage about going to court to fight the final results. In the second topic, viz. the outbreak of protests and violence in the aftermath of the elections, the prominence of the police is striking. Three times *police* features as keyword: this shows that the local press not only focused on violence amongst the Kenyan people but also criticized the role and the actions of state-controlled law enforcement. The appearance of media in the topic boxes refers to the government-issued media ban on live coverage of the disturbances. The third topic (at the right) concerns the efforts to get out of the political deadlock. The local press coverage concentrated on mediation and talks between members of the opposition and the government. Chief mediator Kofi Annan, who in the end succeeded in reconciling the rivaling parties, received most attention. But also other international support from both American (*Bush, Rice*) and African (*Kikwete*)[20] commentators and conciliators is exhaustively covered.

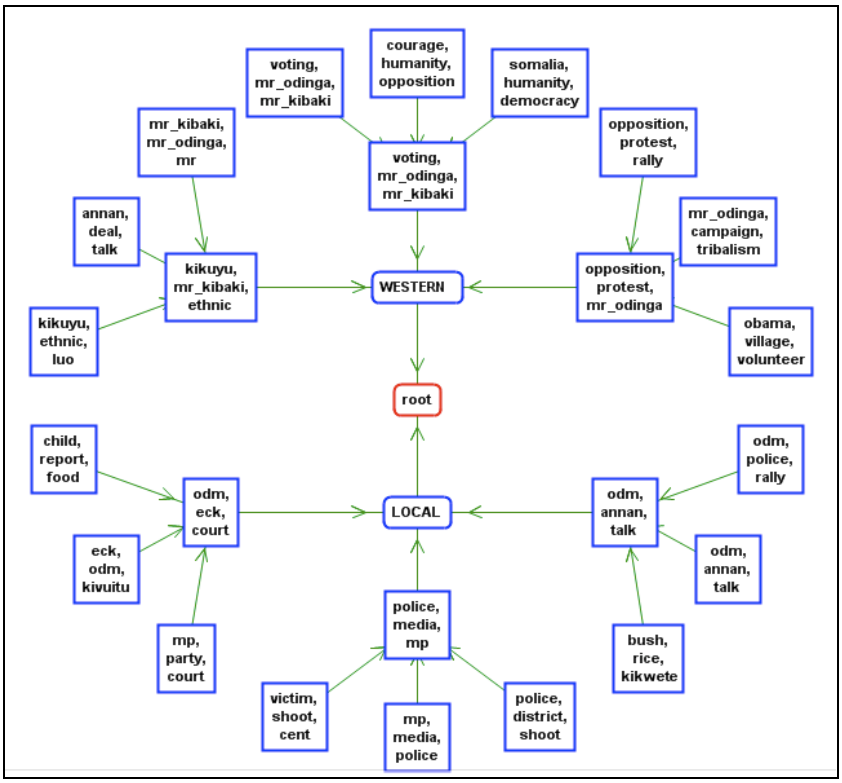

Figure 4: The topic ontology of local and Western press articles on Kenyan elections.

The Western newspapers roughly touch upon the same main topics, but they have different accents. The cluster described by the topics *voting, Mr Odinga, Mr Kibaki*, represents the articles about the voting process. The duel between Kibaki and Odinga is emphasized as well as the humanity of the elections. Furthermore Kenya as 'a stable democracy' is compared to less stable neighbors such as Somalia. The right main topic, typified by the keywords *opposition, protest, Mr Odinga,* deals with the violent escalation of sociopolitical unrest. Here the presence of the keyword *tribalism* in the subtopic is telling. The third main topic concerns the conflict resolution and peace talks. However, not the attempts to find solutions for the political deadlock and the violence in society receive primary attention, but rather the polarization between the disputing parties and the ethnicization of the conflict is focused on. In that way, the

[20] Jakaya Mhriso Kikwete, the president of neighboring country Tanzania, was chairman of the African Union from 31 January 2008 to 2 February 2009.

topic ontology corroborates earlier observations about the tribal frame that is used in the Western press. This issue will be picked up in Section 7.

7. Interpretation from a pragmatic perspective

Contrary to a classical content analysis we subjected our corpus to a number of text mining techniques to get an insight into the data and to procure computational support for further qualitative analysis. By means of a pragmatic analysis, the discovered differences between Western and local news are contextualized and put into perspective. This section presents a summary of our major findings and provides interpretations from a pragmatic point of view.

With the experimental results in mind, pragmatic discourse analysis was guided by the following question: What is the meaning of the lexical choices uncovered by text mining in the contexts of the Western and/or local newspaper reports about the described events? More specifically, how do references to ethnicity, triggered by such words as *tribe, tribalism* or *ethnic*, function in the discourses and concrete contexts in which they occur? These questions must be tackled by a coherent methodology (cf. Verschueren 1996, 1999, 2008).

Without going into specifics, the most important methodological requirements concern the dataset, text analysis and triangulation. Ideally, the corpus should be large enough, so that the discourse analysis pertains to a variety of data types (e.g. different genres, text sources, discourse domains, …) and multiple levels of linguistic structure. Therefore we included both hard news reports, opinion articles, features and editorials. However, in this paper we only focused on patterns of topical, representational and lexical choices, ignoring other grammatical or discursive levels of linguistic structure. Text analysis means close scrutiny of linguistic elements, not only within one and the same text, but also between different texts and discourses. Thus it comprises also intertextual and contextual analysis. Different texts are compared by carefully (re)reading them and interpreting the meanings generated by their language use. By concentrating on contrast and variability, explicit and implicit patterns of meaning can be exposed. Since 'ideological' ideas and beliefs are often carried along implicitly in the discourse, a linguistic pragmatic study of ideology in discourse goes beyond the explicit content in the search for patterns of unquestioned implicit meaning. As triangulation is concerned, the results should be subjected to counter-screening by looking for evidence that would contradict one's research conclusions. A discourse analyst has to take a critical and self-reflective stance towards his own research.

As a starting point for our pragmatic discourse analysis, it should be recognized that at first sight there are more similarities than differences between the local and Western newspaper articles. This might not be so surprising, given that the newspaper reports deal with the same events. A fair amount of intertextuality not only became clear in the linguistic analysis of keywords presented in Section 4.2, but is also evident from the text mining experiments. The topic ontology in Figure 4, for instance, shows that the Western and local press coverage share topical words, such as *Annan* and *talk.* However, the intertextuality cannot only be explained on the basis of a thematic relationship between the various newspaper reports. Other important factors are journalistic practices and commonly shared criteria of news values. In addition, similarities in the reporting and the overlap of meanings in our case study are also due to mutual influence.

Let us briefly explain these claims, because the obviousness of correspondences

makes the numerous subtle differences all the more salient. News workers display universal preferences and intuitions about the nature of news (e.g. Galtung and Ruge 1965; Van Dijk 1988; Bell 1991; Westerståhl and Johansson 1994; Harcup 2004; Pape and Featherstone 2005; Wu 2007; and Machin 2008). When an event is negative, abnormal or unusual, clearly delineated, easily interpretable and explainable for the target audience thanks to readily available sources, it will likely become news. On the contrary, when the event is (perceived to be) hardly deviant from daily reality, when it is the result of long societal processes, or when it requires a lot of historical background knowledge to comprehend, the event will have a low news value. After all, news must fit into the readers' familiar frame of interpretation. Finally, press coverage is also partly affected by other press coverage. Journalists report on what their colleagues report. They not only often go to the same press conferences and share interviewees, but they also read other newspapers and exchange information.

But the interpretations of the same events can be quite diverse in different news markets. The way words are used is crucial. For instance, quantitatively the topic of violence is more or less equally covered, but a qualitative pragmatic analysis combined with text mining reveals that the Western press often puts instances of violence into a tribal frame, either explicitly labeling violence as tribal and ethnic, or linking it to the ethnicity of the perpetrators (see below). In the local press the violence is connected either to political protest or to criminal behavior without any explicit references to the ethnicity of the people involved. Moreover, a high degree of intertextuality does not mean that there are no intertextual gaps. By reading the corpus articles one quickly observes that the Kenyan press provides more varied interpretations of the multiplicity of conflicts than the Western press, although they fail to explain the ethnic factor of certain conflicts. A glimpse at Figure 4 indicates that the keywords generated for the local press show a greater variety than the recurring keywords of the Western media.

Our first finding pertains to the first part of the double hypothesis that followed from the text mining analyses: the Western media covered the news through a dominant ethnic lens. All the presented classification models show the Western media's tendency to link the Kenyan election crisis to ethnic divisions. Different lexical strategies are used to create a tribal frame of interpretation. Firstly, this is done by explicit references to the tribes of the actors involved. The text mining models indicate that an explicit reference to Kibaki's tribe of the Kikuyu is one of the most important distinguishing features of the Western newspaper accounts (see for example Tables 3, 4, 5 and 6, and Figure 2). Although explicit references to other tribes, like the Luo, Kalenjin or Luhya, did not surface as principal classifier features, closer investigation showed that they occur almost exclusively in the Western press. The rare instances of *Luo* or *Kikuyu* in the Kenyan press all referred to regions or political entities as in *Luo Nyanza*, referring to the part of the Nyanza province that is inhabited by the Luo, or as in *Kikuyu parliamentary seat*. The latter is an example of the ethnicization of Kenyan politics, which is often taken for granted in the Kenyan press. Compare this specific use of tribe names to the often generalizing and stereotyping way in which the ethnic communities are named in the Western press:

(1) The election has uncorked dangerous resentment toward the Kikuyus, the privileged ethnic group of Kenya, who have dominated business and politics since independence in 1963.
(*New York Times*_ Fighting Intensifies After Election in Kenya_01/01/2008)

(2) More than 200 people, mainly Kikuyus, the same tribe as President Mwai Kibaki, were sheltering for safety in the Kenya Assemblies of God church five miles outside Eldoret in the Rift Valley. An armed gang of young men drawn from the Kalenjin, Luhya and Luo tribes ethnic groups [sic] which backed the beaten presidential candidate Raila Odinga stormed the church compound yesterday morning and set it alight.
(*The Independent*_80 children massacred in Kenyan church_02/01/2008)

(3) Unconfirmed reports said that gangs of Kikuyu youths had hunted down Luos, stripped them naked and forcibly circumcised them.
(*The Times*_Kenya teeters on the brink_03/01/2008)

These examples also illustrate other ways of ethnicizing the news coverage: the use of the ideologically marked word pairs *tribe* and *tribal*, *ethnic group* and *ethnic*.

Especially in the 'Western world', the noun *tribe* tends to carry the negative connotation of primitiveness and savageness (Krishnamurthy 1996). It is a value-laden term, what McGee would call an 'ideograph', i.e. a keyword of the discourse that functions as a building block of ideology, signifying a unique ideological commitment, and which not only "warrants the use of power, excuses behavior and belief [but also] guides behavior and belief into channels easily recognized by a community as acceptable" (McGee 1980: 15). The occurrence of *tribe* in Table 1 is a first indication that it is prominent in the Western press coverage. This observation is corroborated by the PRISM rules in Table 5 and by the contrasting keywords analysis in Table 6.

Also the adjective *tribal* was often selected in the text mining processes as a distinctive feature. The word *tribal* appeared 182 times in 100 different Western articles, compared to its presence in only 14 articles of the local media. A more detailed discourse analysis, in which the terms are examined in their contexts of use, reveals that *tribe* and *tribal* are usually employed in negative contexts, viz. contexts of physical and verbal violence, of corrupt politics and courts, and of other crisis situations, hence its typical negative connotation in the Western media. In more positive contexts like those of health care or human rights ethnicity is not brought to the fore. For instance, in the *Independent* article 'A chilling tour of the Kenyan church that became the scene of mass murder' (3 January 2008) the attackers and the attacked are pinned down on their ethnicity. Also the tribal affiliations of the political leaders are always explicitly mentioned, while the tribe of the deputy director of the hospital is not given. Likewise the tribe of the often quoted chairman of the Kenya National Commission on Human Rights can only be guessed. If the words *tribe* and *tribal* are used by the Kenyan press, they often occur in contexts of disapproval or denial. The contexts of these words involve places or political entities, such as voting blocs. While in the Western press tribalism is the main explanation of most conflicts throughout Kenya, the conflicts are clearly localized in the Kenyan media and the tribal factor is sometimes even explicitly denied, as in (4) and (5).

(4) ODM said the mayhem is not an expression of tribal hate but citizens' cry for their democratic rights.
(*The Standard*_ODM: Chaos is citizen's demands for their rights_04/01/2008)

(5) Unlike in the previous elections where tribalism was the main factor, the ongoing insecurity being witnessed in the [North Rift] region is more of land

> politics [sic].
> (*The Saturday Nation*_The land factor in violence that has rocked North Rift_05/01/2008)

Only once is the violence referred to as *tribal clashes* in our Kenyan sample. In this case, the article 'Pope calls for end to violence' in the *The Sunday Standard* of 6 January 2008 appeared to be literally taken from the Reuters press agency. In general terms, the violence gets the qualification *post-election* in the Kenyan press (see Prism rule 6 for class LO in Table 5). Otherwise, concrete instances of violence are specified and again clearly spatio-temporally situated. Note, however, that even clear instances of ethnic violence are never specified as such in the local press, where the ethnic factor of certain conflicts tends to be obscured (see below).

The use of *ethnic group* and *ethnic* may be less ideologically marked and more politically correct than *tribe* and *tribal.* However, both pairs often seem to be used interchangeably, as the error in example (2) suggests (see also Krishnamurty 1996: 132). In the topic ontology generated with OntoGen (Figure 4), the adjective *ethnic* appears as topic descriptor of two document clusters, which makes it a key topic of the Western news content. Although *ethnic group* does not have the same connotation as *tribe* and the word *ethnic* is sometimes used in more neutral contexts (e.g. when there is reference to ethnic areas, ethnic communities, ethnic neighbors, ethnic solidarity), the label ethnic is still often applied to negative nouns such as *conflict, violence, fighting*, and *tensions*. We also found that the adverb *ethnically* only appears in the Western media (e.g. in the expression *ethnically charged violence*), 36 times to be precise.

Our next major finding concerns the second part of our double hypothesis. Contrary to the Western news texts, the tribal factor is downplayed in the local press coverage. Several models confirmed that the Kenyan reporting is typified by a sociopolitical perspective. The series of conflicts is seen as the consequence of the elections, vote-rigging, political incitement. When described in general, all these conflicts are labeled in political terms as a *post-election crisis*, a *political impasse*, a *political stalemate* or as a *humanitarian crisis*. Perpetrators of violence and their victims are not named by their tribe. Rather they are presented as supporters of political parties or as unspecified *youths*, *gangs*, *mobs*, *protesters* or *criminals*, even when clashes clearly have an ethnic aspect. The murdering of Kikuyus in the Kenya Assemblies of God church (see also example (2)) is reported in the Kenyan newspapers without any references to tribes:

(6) This came on a day the post-election violence that has rocked parts of the country took serious proportions when at least 30 children and 10 adults who had sought refuge in a church were burnt to death in acts of violence linked to protests against the President's re-election.
(*The Standard*_Peace calls amid continued bloodletting_02/01/2008)

(7) According to those who escaped the killings, they have never had a problem with the community they have lived with in the village for the last 40 years and the attack caught them by surprise.
(*Daily Nation*_Raid on displaced families that shocked the world_06/01/2008)

The Kenyan newspapers use the word *community*, which has a weaker, more neutral connotation, instead of *tribe* or *ethnic group*. Note the political framing of the events in example (6). A link is created with political affairs, while in (7) the events are illuminated from a social point of view. But the ethnic dimension of this particular

conflict is ignored.

In Figure 4, political terms dominate the local topic ontology (e.g. *ODM, ECK, MP*, *party, rally, (mediation) talk).* One of the most distinguishing features for the local class of newspapers is the abbreviation of Odinga's party, ODM. From several models as well as from the SVM keyword list we can infer that the local media preferred political party names above references to ethnic groups. For instance the JRip models in Tables 3 and 4 show that ODM is a typical feature of the local press and appears to stand in contrast to the word *opposition*: in both models, the first (thus most important) rules contain a combination of these two words (presence of *ODM* and absence of *opposition* in the local media and vice versa in the Western media). Table 4 also makes clear that instead of describing the troubles as a *tribal struggle*, *ethnic fighting* or *civil war*, the Kenyan press rather relates the events to the political dispute or a broader sociopolitical *crisis*, which is presumed to have triggered them (see Rules 4 and 5).

The framing of the events affects how the main participants of the news stories are presented. Our third finding relates to the different representation of the main protagonists. To continue our examination of the functioning of the word *opposition,* we started to explore the context in which this word is used. Looking at the use of lexical choices in context we discovered that *opposition* is often used in the expression *opposition leader*, referring to Raila Odinga (e.g. in phrases such as "to enter talks with the opposition leader Raila Odinga" or "President Mwai Kibaki and the opposition leader Raila Odinga"). Alternatively, in the local press Raila Odinga is represented as a presidential candidate or party member of ODM, proof of which can be found in the J48 decision tree in Figure 3. Odinga is referred to as *Mr Raila Odinga*, frequently complemented by the genitive modifier of *ODM*, or as *a leader of ODM.* Further close reading detected frequent uses of *ODM's Raila Odinga* and *the ODM presidential candidate*.

Of course, the manner in which people are described depends on the background knowledge of the target audience. For a Western audience the information of Odinga being *opposition leader* is more relevant than his characterization as a member of ODM. On the other hand, the specification of the party is more informative for a Kenyan readership. Although it is inaccurate and a gross simplification of Kenyan politics to call Odinga *the* leader of *the* opposition, it can be explained by the notion of domestication. As Lee et al. (2000) concluded, international news is often adapted to local frames of interpretation. When the British reader is used to a political system in which there is an opposition and a ruling party, the journalist might decide to conceptualize the foreign political system likewise, so that the newspaper report can easily be understood. Similarly, a Kenyan journalist does not have to state explicitly the ethnicity of the actors as his readers know what tribe Odinga and Kibaki are from. Nevertheless, it is striking how the Western media portrayed the protagonists, constantly emphasizing their ethnic origin. Extra information is usually given in non-restrictive relative clauses, placed between commas. That is why in Figure 3 *Mwai Kibaki* and *Raila Odinga* followed by a comma came out as identifying trigrams for the Western press. Note the contrast between (8) and (9).

(8) Mr Odinga, the son of Jaramogi Oginga Odinga, the trade-unionist independence hero and first vice-president of Kenya, was educated in East Germany and called his first son Fidel. Like all Kenyan politicians he is a wealthy businessman and dropped the socialist rhetoric long ago. Nevertheless, as a Luo from the poor Lake Victoria region of Western Kenya, he appeals to

marginalised communities much more than the elitist Mr Kibaki, who is a Kikuyu.
(*The Times*_Democracy comes out fighting as Kenyan voters take off the gloves_27/12/2007)

(9) The impact of the logistical problems was felt in Lang'ata constituency where ODM presidential candidate and former local MP, Mr Raila Odinga's name was among thousands missing from the polling register.
(*The Standard*_Kenyans make huge statement_28/12/2007)

By means of the repeated use of specific representations, such as in example (8), or as in the *New York Times* article 'Kenyans vote in test of democracy' (28/12/2007), in which Kibaki is introduced as "a courtly gentleman and economics whiz" but also as "a tribal politician" and Odinga as "a rich, flamboyant businessman who rides around in a bright red $100,000 Hummer", the Kenyan politicians become caricatures.

8. Discussion and further work

This paper provides a methodological example of a fruitful collaboration of two, traditionally separated, methodological frameworks and illustrates this methodology through a specific case study of reporting on the aftermath of the Kenyan elections. In this section we evaluate the tested text mining techniques, critically reflect on their usefulness for pragmatic research and summarize some of the new insights into the discourse under study.

We hope to have shown that quantitative computational methods, more specifically the text mining methods, and qualitative pragmatic language research are not irreconcilable. What is more, they are complementary and enable better insights into the subject of investigation. In our case the combination of text mining and linguistic-pragmatic analysis constitutes a critical news discourse analysis with special attention to ideological differences in national versus foreign press coverage. The link with ideology is established by the fact that most of the differences we found usually remain hidden for the readers. They are unquestioned and taken for granted.

The text mining methods have focused on discovering contrasting views of national, Kenyan (called *local*), and foreign, British and American (called *Western*), press coverage. The text mining approaches used were all based on generating interpretable information: either by predictive classification models (decision trees and decision rules models) or by using descriptive topic ontology text mining approaches combined by contrasting keywords detection. Linguistic pragmatics, as elaborated by Verschueren (1996, 1999, 2008), was used as an interpretative methodology. The observed lexical choices follow from a natural way of seeing things, informed by an underlying ideology on the basis of which journalists try to make sense of the world. A system of commonsensical, normative ideas and beliefs, partly shaped by and adapted to the context-specific events and the concrete circumstances in which they had to operate, caused Kenyan journalists to steer clear from tribal references. For the Western press, we noticed that the "[m]edia tie their narratives selectively to larger historical frameworks to achieve interpretative coherence" (Lee et al. 2000: 307). Because well-known ethnic conflicts raged in the past through different places in Africa and because tribal affiliations do matter in Africa, (new) conflicts are easily interpreted in ethnic terms.

Our text mining experiments indicated that the major difference between the Western and the local press lies in the framing. A tribal frame was created in the British and American newspapers, while the Kenyan dailies opted for a sociopolitical frame. This was done by the Western media when they explained the conflicts in Kenya mainly as tribal animosity or struggles for power between tribes, or when they compared the post-election crisis to the Rwandan genocide. No doubt a lot of the conflicts in the aftermath of the 2007 general election did have an ethnic aspect, but always the conflicts were more complex. In a multi-ethnic country like Kenya tribe is part of people's identity, but the question can be asked whether it is always relevant to emphasize the tribe when introducing people into news discourse. As Ray (2008: 8) cogently argues, the "widespread and reckless usage of the term 'tribe' and its various permutations hinders the ability of readers to understand how ethnic identities have evolved and interacted with one another in Kenya over time, and in relation to such factors as state and class formation; economic, social and political change; as well as more mundane facts of life such as migration and intermarriage". Reducing the interpretation of the conflicts to tribal clashes is a choice that prevents accurate understanding of the causes of these conflicts which in most cases also had considerable political, social and economic dimensions. When *The New York Times* writes that "the election seems to have tapped into an atavistic vein of tribal tension" ('Disputed vote plunges Kenya into bloodshed', 31 December 2007), a complex political and social phenomenon is reduced to primordial sentiments of unchanged and unchanging gangs of opposing tribes, while Ray (2008: 9) observes that "where ethnicity has played a role in post-independence violence, it is not because of ancient hatred but rather because of a perceived relationship between ethnicity and access to material resources and political power, which has its roots in the 20th century".

By their strategies of generalization and simplification in focusing on tribes and tribal violence the Western media provided a rather one-sided view of the complex reality. Ogola (2009: 62) rightly remarks that "inequitable allocation of resources, the failure to undertake comprehensive constitutional reforms, the monopolization of the political process by the elite, the arbitrary exercise of state power and the normalization of the state and its various institutions together provided conditions for political instability which ultimately contributed to the 2007 election crisis". These aspects were dealt with in the Kenyan press.

The quality of the local newspaper reporting, however, is not without discussion either. Rambaud (2008: 77) contends that the Kenyan "print media treated the elections in a balanced and responsible manner", avoiding a reduction to ethnic-only explanations. He acknowledges that ethnicity was a regular theme in the press but particularly in the opinion pages, adding that "*The Daily Nation* and *The Standard* denounced its overhyping" (Rambaud 2008: 82). Ogola (2009) on the other hand criticizes the Kenyan press for covering up instead of covering accurately and impartially. He even claims that "the deliberate deletion of ethnic references in stories merely helped reify the news media's framing of the conflict as unambiguously ethnic" (Ogola 2009: 59). By applying self-censorship and anxiously avoiding the ethnic factors that were for many Kenyans very obvious in some conflicts, the issue of ethnicity became conspicuous in its absence in the Kenyan press. Although our observation is that the reporting in both *The Standard* and *Daily Nation* was more balanced and varied than that in newspapers from the US or UK, the question remains why the local press at all cost shied away from references to tribe. By doing so they also failed to provide an accurate account of the events. Although they illuminated the events from different angles, they tended to neglect the ethnic perspective, even when

it did play a role.

Coming back to our methodological collaboration between text mining and pragmatics, the findings presented in this paper could not be reached purely on the basis of text mining results, nor by pragmatic analysis alone. When applying text mining, one should be aware of the limitation that simple text mining approaches are incapable of analyzing the words in their context and ignore many important phenomena, such as negation, modality or the impact of headlines, crossheads and other news discourse specific text parts. Therefore, text mining results become really significant and socially relevant when they are reconnected with context in combination with a linguistic-pragmatic analysis in which the functioning of structural and linguistic choices in their contexts of use is examined. A pragmatic analysis takes co-text and context into account, not only by examining intertextual or interdiscursive relations, but also by appealing to related discourses from other contexts, such as election and human rights reports, and by taking the broader socio-cultural climate into consideration, which is essential to understand the automatically generated text mining results.

As manual linguistic-pragmatic analysis can be very laborious, especially when working with large corpora, text mining, as proposed in our framework, can be helpful in three ways. Firstly, it can be useful to get an initial, orienting view of a large-scale corpus. Secondly, it can indicate the further direction of the linguistic-pragmatic analysis by automatically pointing out contrasting words and patterns (for building new hypotheses and discovering new knowledge). Thirdly, text mining results can be used during the counter-screening phase. The quantitative results can be employed to check whether or not there are implications or patterns that contradict the research conclusions. On the other hand, as a limitation of our methodology, text mining tools are not yet sufficiently user friendly for direct use by linguists and much of text preprocessing needs to be done before using the tools.

To conclude, in our approach, text mining results are put back into the context and interpreted by pragmatic analysis, while pragmatic analysis benefits from the patterns and models built independently from a researcher's subjective view so that they help to avoid the risk of reading more in(to) the text than is warranted. That is why we see text mining as a useful methodology in support of further pragmatic analysis. Moreover, we do not see text mining as subservient to pragmatic text analysis. Rather, we suggest an inclusive methodology of pragmatic discourse analysis in which text mining is combined with linguistic-pragmatic analysis so as to gain a deeper insight into the discourses under investigation and in order to deliver accurate interpretations without falling into the traps of underinterpretation or overinterpretation (O'Halloran and Coffin 2004).

In future work, we plan to further develop our framework which combines text mining and pragmatic analysis, and extend it to cross-lingual domain modeling which would give the possibility of performing a similar analysis on comparable corpora of different languages. An important aspect of future analysis will be the consideration of syntactic patterns (Luyckx and Daelemans 2008), modality structures and reported speech (see the emerging field of sentiment analysis and opinion mining, e.g. Liu 2010.

Acknowledgements

The work of the authors was supported by the funding of their national research projects, supplemented by a study grant from the Flemish Government which supported the work of the first author. The corpus was collected as part of the *Intertextuality and Flows of Information* project of the IPrA Research Center (University of Antwerp). We are grateful to Kim Luyckx for her help in data preprocessing and Guy de Pauw and Jef Verschueren for their interest and support in the preparation of this work. Finally, we are very grateful also to Richard Wheeler and to two anonymous reviewers whose comments enabled significant improvements of this paper.

References

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

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyzanowski, T. McEnery, and R. Wodak (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: 273–306.

Balahur, A., and R. Steinberger (2009) Rethinking sentiment analysis in the news: From theory to practice and back. In *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis*, Satellite to CAEPIA 2009.

Bell, A. (1991) *The Language of News Media*. Oxford: Blackwell.

Cendrowska, J. (1987) PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies* 27.4: 349–370.

Cohen, W. (1995) Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, p. 115–123.

Cohen, W., and Y. Singer (1999) Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems (TOIS)* 17.2: 141–173.

Daelemans, W., S. Bucholz, and J. Veenstra (1999) Memory-based shallow parsing. In *Proceedings of the Computational Natural Language Learning Workshop* (CoNLL-99). Demo: http://www.cnts.ua.ac.be/cgi-bin/jmeyhi/MBSP-instant-webdemo.cgi

EU EOM Kenya (2008) Kenya: Final Report. General Elections 27 December 2007 (3 April 2008). Brussel: EU EOM Kenya, retrieved from http://www.eueom.eu/ [01/03/2010].

Fairclough, N. (1995). Media Discourse. London: Arnold.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996) The KDD process for extracting useful knowledge from volumes of data. *Communication of the ACM* 39. 11: 27–34.

Feldman, R., and J. Sanger (2007) *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Fielding, N.G., and R.M. Lee (1998) *Computer Analysis of Qualitative Research.* London: Sage.

Finn, A., and N. Kushmerick (2006) Learning to classify documents according to genre. In *Journal of the American Society for Information Science and Technology* 57.11: 1506–1518.

Fortuna, B., C. Galleguillos, and N. Cristianini (2009) Detecting the bias in media with statistical learning methods. In N. Ashok, Srivastava and M. Saham (eds.), *Text Mining: Theory and Applications*. London: Taylor and Francis Publisher.

Fortuna, B., M. Grobelnik, and D. Mladenić (2006) System for semi-automatic ontology construction. In *Proceedings of the Demo Session at European Semantic Web Conference ESWC* (2006).

Fortuna, B., M. Grobelnik, and D. Mladenić (2007) OntoGen: Semi-automatic ontology editor. In M.J. Smith, and G. Salvendy (eds.), *Proceedings of Human Interface, Part II, HCI International 2007*, LNCS 4558, Springer, p. 309–318.

Galtung, J., and M.H. Ruge (1965) The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research* 2.1: 64–91.

Gibbs, G.R. (2004) Computer-assisted Qualitative Data Analysis (CAQDAS). In M.S. Lewis-Beck, A. Bryman, and T.F. Liao (eds.), *The Sage Encyclopedia of Social Science Research Methods* (1). Thousand Oaks: Sage, p. 87–89.

Greevy, E.P., and A.F. Smeaton (2004) Text categorisation of racist texts using a support vector machine. In *Proceedings of 7es Journées internationales d'Analyse statistique des Données Textuelles JADT (1)*. Leuven: PUL, p. 533–544.

Harcup, T. (2004) *Journalism: Principles and Practice*. London: Sage.

Harris, R.J. (2004) *A Cognitive Psychology of Mass Communication* (4th ed.) Mahwah: Lawrence Erlbaum.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman.

Koller, V., and G. Mautner (2004) Computer applications in critical discourse analysis. In C. Coffin, A. Hewings, and K. O'Halloran (eds.), *Applying English Grammar: Functional and Corpus Approaches*. London: Arnold, p. 216–228.

Krishnamurty, R. (1996) Ethnic, racial and tribal: The language of racism? In C.R. Caldas-Coulthard, and M. Coulthard (eds.), *Texts and Practices: Readings in Critical Discourse Analysis*. London/New York: Routledge, p. 129–149.

Lee, C., J.M. Chan, Z. Pan, and C.Y.K. So (2000) National prisms of a global 'Media Event'. In J. Curran, and M. Gurevitch (eds.), *Mass Media and Society* (3rd ed.). London: Arnold., p. 295–309.

Lin, W.-H., E. Xing, and A. Hauptmann (2008) A joint topic and perspective model for ideological discourse. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, p. 17–32.

Lindlof, T.R., and B.C. Taylor (2011) *Qualitative Communication Research Methods* (3rd ed.). Thousand Oaks: Sage.

Liu, S.-Z., and H.-P. Hu (2007) Text classification using sentential frequent item sets. In *Journal of Computer Science and Technology* 22.2. Beijing: Institute of Computing Technology, p. 334–337.

Liu, B. (2010) Sentiment Analysis: A Multi-Faceted Problem. *IEEE Intelligent Systems* 25.3.

Lüdeling, A., and M. Kytö (eds.) (2008) *Corpus Linguistics. An International Handbook.* Berlin: Mouton de Gruyter.

Luyckx, K. (2010) *Scalability Issues in Authorship Attribution*. Brussels: UPA University Press Antwerp.

Luyckx, K., and W. Daelemans (2008) Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics*

(COLING 2008), p. 513–520.

Machin, D. (2008) News discourse I: Understanding the social goings-on behind news texts. In A. Mayr (ed.), *Language and Power: An Introduction to Institutional Discourse.* London: Continuum, p. 62–89.

MacMillan, K. (2005) More than just coding? Evaluating CAQDAS in a discourse analysis of news texts. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6.3, art. 25.

Mahlberg, M. (2007) Lexical items in discourse: Identifying local textual functions of sustainable development. In M. Hoey, M. Mahlberg, M. Stubbs, and W. Teubert (eds.), *Text, Discourse and Corpora. Theory and Analysis*. London/New York: Continuum, p. 191–218.

Matu, P.M., and H.J. Lubbe (2007) Investigating language and ideology: A presentation of the ideological square and transitivity in the editorials of three Kenyan newspapers. *Journal of Language and Politics* 6.3: 401–418.

Mautner, G. (2007) Mining large corpora for social information: The case of elderly. *Language in Society* 36.1: 51–72.

McGee, M.C. (1980) The 'ideograph': A link between rhetoric and ideology. *The Quarterly Journal of Speech* 66.1: 1–16.

Mitchell, T. (1997) *Machine Learning*. Boston: McGraw Hill.

Morley, J., and P. Bayley (2009) *Corpus-Assisted Discourse Studies on the Iraq Conflict: Wording the War*. New York: Routledge.

Ngonyani, D. (2000) Tools of deception: Media coverage of student protests in Tanzania. *Nordic Journal of African Studies* 9.2: 22–48.

Ogola, G. (2009) Media at cross-roads: Reflections on the Kenyan news media and the coverage of the 2007 political crisis. *Africa Insight* 39.1: 58–71.

O'Halloran, K. (2010) How to use corpus linguistics in the study of media discourse. In A. O'Keeffe, and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics.* London/New York: Routledge, p. 563–577.

O'Halloran, K., and C. Coffin (2004) Checking overinterpretation and underinterpretation: Help from corpora in critical linguistics. In C. Coffin, A. Hewings, and K. O'Halloran (eds.), *Applying English Grammar: Functional and Corpus Approaches*. London: Arnold, p. 275–297.

O'Keeffe, A., B. Clancy, and S. Adolphs (2011) *Introducing Pragmatics in Use.* London: Routledge.

Oloo, A.G.R. (2007) The contemporary opposition in Kenya: Between internal traits and state manipulation. In G.R. Murunga, and S.W. Nasong'o (eds.), *Kenya: The Struggle for Democracy*. Dakar: CODESRIA Books, p. 90–125.

Pape, S., and S. Featherstone (2005) *Newspaper Journalism: A Practical Introduction*. London: Sage.

Quinlan, J. (1993) *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann.

Rambaud, B. (2008) Caught between information and condemnation: The Kenyan media in the electoral campaigns of December 2007. In J. Lafargue (ed.), *The General Elections in Kenya, 2007 (Special issue of Les Cahiers d'Afrique de l'Est* (38)). Nairobi: IFRA, p. 57–107.

Ray, C. (2008) How the word 'tribe' stereotypes Africa. *New African* 471: 8–9.

Reah, D. (1998) *The Language of Newspapers*. London/New York: Routledge.

Richardson, J.E. (2007) *Analysing Newspapers: An Approach from Critical Discourse Analysis*.

Basingstoke: Palgrave Macmillan.

Rühlemann, C. (2010) What can a corpus tell us about pragmatics? In A. O'Keeffe, and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge, p. 288–301.

Scott, M. (2008) WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

Schönfelder, W. (2011) CAQDAS and qualitative syllogism logic—NVivo 8 and MAXQDA 10 Compared [91 paragraphs]. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 12(1), art. 21.

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys* 34.1: 1–47.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2000) Automatic text categorization in terms of genre and author. *Computational Linguistics* 26.4: 471–495.

Stubbs, M. (1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. (2001) Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics* 22.2: 149–172.

Thornbury, S. (2010) What can a corpus tell us about discourse? In A. O'Keeffe, and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge, p. 270–287.

Van Dijk, T.A. (1988) *News as Discourse*. Hillsdale: Lawrence Erlbaum.

Van Dijk, T.A. (2006) Ideology and discourse analysis. *Journal of Political Ideologies* 11.2: 115–140.

Van Ginneken, J. (2002) *De schepping van de wereld in het nieuws: De 101 vertekeningen die elk 1 procent verschil maken* (2nd ed.). Kluwer: Alphen aan den Rijn.

Van Leeuwen, T. (2008) *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford: Oxford University Press.

Verschueren, J. (1996) Contrastive ideology research: Aspects of a pragmatic methodology. *Language Sciences* 18.3/4: 589–603.

Verschueren, J. (1999) *Understanding Pragmatics*. London: Arnold.

Verschueren, J. (2008) Context and structure in a theory of pragmatics. *Studies of Pragmatics* 10: 13–23.

Westerståhl, J., and F. Johansson (1994) Foreign news: News values and ideologies. *European Journal of Communication* 9: 71–89.

Witten, I.H., and E. Frank (2005) *Data Mining Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Elsevier.

Wrong, M. (2008) Don't mention the war. *New Statesman* 137.4884: 22–23.

Wu, D.H. (2007) A brave new world for international news? Exploring the determinants of the coverage of foreign nations on US websites. *The International* Communication Gazette 69.6: 539–551.

Zhao, Y., and J. Zobel (2005) Effective and scalable authorship attribution using function words, *LNCS* 3689, p. 174–189. Berlin/Heidelberg: Springer.

ROEL COESEMANS is a Doctoral Candidate in Linguistics at the IPrA Research Centre, University of Antwerp. He is currently working towards the completion of his Ph.D. dissertation about implicit meanings and ideology in national as compared to international newspaper reports, covering events set in Africa. This ethnographically-supported news discourse analysis from a pragmatic perspective is part of the *Intertextuality and Flows of Information* research project, in which processes of meaning generation and transformation in international newspaper reporting are studied. His main research interests include journalism, media representation and pragmatic tools for contrastive media analysis, such as presupposition and implicature.
Address: IPrA Research Center, University of Antwerp, Antwerp, Belgium. E-mail: roel.coesemans@ua.ac.be

WALTER DAELEMANS is research director of CLiPS, the Computational Linguistics and Psycholinguistics research centre of the department of linguistics at the University of Antwerp. His main research interests are in machine learning of language, computational psycholinguistics, stylometry, and text mining. He has coordinated or participated in several national and European projects on text mining and computational linguistics, and is (co-)author of several publications in these areas, among others of a monograph on Memory-Based Language Processing with Cambridge University Press (2005). Until now, sixteen Ph.Ds have successfully obtained their degree under his (co-)supervision.
Address: CLiPS – Computational Linguistics Group, University of Antwerp, Antwerp, Belgium.
E-mail: walter.daelemans@ua.ac.be

NADA LAVRAČ is Head of the Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia, and Professor at the University of Nova Gorcia in Slovenia. Her main research interests are in machine learning, relational data mining, text mining, knowledge management, and applications in medicine and bioinformatics. She was the scientific coordinator of the European Scientific Network in Inductive Logic Programming (ILPNET, 1993-1996) and co-coordinator of the 5 FP EU project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (SolEuNet, 2000-2003). She is author and editor of several books, including *Inductive Logic Programming: Techniques and Applications* (Kluwer 1997), and *Relational Data Mining* (Springer 2002), and *Foundations of Rule Learning* (Springer 2012).
Address: Jožef Stefan Institute, Ljubljana, Slovenia; University of Nova Gorica, Nova Gorica, Slovenia. E-mail: nada.lavrac@ijs.si

SENJA POLLAK is junior researcher at the Department of Translation Studies, Faculty of Arts, University of Ljubljana, Slovenia. After the BSc in French linguistics (Sorbonne 3, Paris) and BSc in Sociology of culture and French language and literature (University of Ljubljana), she oriented her research into Computational Linguistics, focusing on text mining. She obtained the Advanced Master in Linguistics (Interdisciplinary Linguistics) MSc degree from the University of Antwerp. Her current research interests as Ph.D. student are multilingual terminology and definition extraction.
Address: Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia. E-mail: senja.pollak@ff.uni-lj.si