Private State in Public Media: Subjectivity in French Traditional and Online News

A. Küppers, Lydia-Mai Ho-Dac

To cite this version:

A. Küppers, Lydia-Mai Ho-Dac. Private State in Public Media: Subjectivity in French Traditional and Online News. 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ECAI 2010, Aug 2010, Lisbon, Portugal. pp.22-29. hal-00983040

HAL Id: hal-00983040
https://hal.science/hal-00983040
Submitted on 24 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Private State in Public Media
Subjectivity in French Traditional and On-line News

Anne Küppers and Lydia-Mai Ho-Dac

Abstract. This paper reports on ongoing work dealing with the linguistic impact of putting the news on-line. In this framework, we investigate differences in one traditional newspaper and two forms of alternative on-line media with respect to the expression of authorial stance. Our research is based on a comparable large-scale corpus of articles published on the websites of the three respective media and aims at answering the question to what extent the presence of the author varies in the different media.

1. Is it a matter of amount and mode of the author’s presence?
2. Is it a matter of lexical choice and diversity?
3. If this were the case, what expressions are used in the respective media?

Our endeavour will be a methodological one. We firstly present our data, and thus describe the different news media included in our study, and the diverse computer aided and manual production steps we performed in order to build up the corpus. Secondly, we outline our working hypotheses that are linked to the chosen types of media and describe the theoretical framework within which they are situated. Thirdly, we present our research method as well as some first results and insights gained throughout the pilot study of our data.

1 Corpus

The main objective of our research is to contrast traditional newspaper language with the language used in alternative forms of journalism in order to determine whether we have to do with distinct genres, or merely different text types.

We therefore created a large-scale corpus consisting of articles published in one traditional newspaper and in two alternative written on-line mass media. Texts included in the corpus have been published between 2005 and 2009 and were collected directly from the respective website’s archives. The media are briefly presented in the following and table 1 gives an overview of the sections included in each data set. Sections are chosen on the basis of comparison of the topics dealt with in order to ensure a higher degree of comparability of the different sub corpora.

The first data set consists of articles published in one of the principal Belgian, French-speaking traditional reference newspapers, namely Le Soir. This liberal, most read supra-regional newspaper was first published in 1887. Texts included in our corpus have been published in the printed or the on-line version of the newspaper.

The second data set is composed of articles published on the website of the French independent journalism project Rue89. This project started in 2007 and aims at unifying professional journalism and Internet culture. It works with a committee of professional journalists and young reporters ensuring a good portion of articles and the reviewing of texts submitted by external domain specialists.

The third data set is made up of articles published on the French alternative on-line information platform AgoraVox. This citizen press website was created in 2005 and follows the principle of editorial democracy, that is to say that any Internet user can subscribe and contribute articles. Reviews are done by members of the committee, composed of some anchormen and the sub-editors, who are potentially all members having published at least 4 articles on the AgoraVox website.

2 Data Processing

The creation of the corpus was realised through several steps. Processing procedures are different for the three media because the constitution of their on-line archives is not uniform. And, as there are no ready-made programs for automatically extracting articles from different web pages ([44]), the data collection was a first challenge. We will briefly describe the basic legs that were realised for the entire corpus.

In a first step, all articles are collected directly from the websites’ archives by means of Perl scripts. The script extracts the list of articles published under a certain URL and saves them in a separate folder for each medium. A second script extracts the articles mentioned in these folders and converts the original files of different source codes into XML format. During this step, the script keeps track of information such as the medium, the section, the author, the title and the date of publication, whenever these are available. Besides, we already partially remove wire copies from our data sets, as they would falsify our analyses on the expression of authorial stance.

For the same reason, interviews, poems and songs are removed manually.

Table 1. Newspaper Sections included in the Data Sets

<table>
<thead>
<tr>
<th></th>
<th>Le Soir</th>
<th>Rue89</th>
<th>AgoraVox</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>World News</td>
<td>International News</td>
<td>Europe</td>
</tr>
<tr>
<td>Politics</td>
<td>Politics</td>
<td>Society</td>
<td>Society</td>
</tr>
<tr>
<td>Culture</td>
<td>Culture</td>
<td>Media</td>
<td>Religion</td>
</tr>
<tr>
<td>Media</td>
<td>Religion</td>
<td>Bizarre</td>
<td>Life and Style</td>
</tr>
<tr>
<td>Society</td>
<td>Bizarre</td>
<td>People</td>
<td></td>
</tr>
</tbody>
</table>

1 Université catholique de Louvain (UCL), Institut Langage et Communication (IL&C), Belgium, email: anne.kueppers@uclouvain.be, lydia.ho-dac@uclouvain.be
3 Hypotheses

On the basis of this multi-layered corpus of written mass media, we aim at bringing to light similarities and divergences in terms of linguistic and structural dimensions. Concerning the language use in the three different media, we have the following hypotheses:

1. Subjectivity is expressed differently in the three media, namely with respect to the amount and lexical choice of subjective expressions.

2. Articles in the alternative media are more subjective than those published in the traditional newspaper, as the author expresses more overtly her/his opinion and thoughts in the former ones.

3. On average, the writing style in articles published in AgoraVox is even more subjective than the style in those published in Rue89.

We formulate these hypotheses on the basis of the working processes prevailing in the editorial departments of the different media and their respective 'philosophy'. The traditional newspaper’s data set is composed of two different types of articles, those published in the paper version and those published on the website. While most of the articles recorded for the printed version are based on investigation and are entirely written by one journalist, most of the articles published on the website are slightly modified wire copies that are simply adapted to the editorial line ([20]). This is a consequence of the prevailing working guideline for the web journalists: 'put the news on-line as fast as possible, be the first to publish, and get the scoop'.\(^2\) This principle influences the writing style, which we assume to be less individual and thus less subjective. As a consequence, the presence of the on-line articles in our data set will probably have an impact on the amount of subjective expressions, their mode and lexical choice.

The two alternative media included in our research differ from the traditional newspaper in some points: First, these media are solely published on the Internet. Second, they work with professional journalists, domain specialists and non-professional Internet users. Third, they tend to cover current topics, but not to get the scoop at all costs. Fourth, the alternative media do not intend to cover all actual topics, as the author expresses more overtly her/his opinion and thoughts in the former ones. This principle influences the writing style, which we assume to be less individual and thus less subjective. As a consequence, the presence of the on-line articles in our data set will probably have an impact on the amount of subjective expressions, their mode and lexical choice.

Concerning their working mode and philosophy, Rue89 defines itself as being more comparable to a radio station than to a traditional newspaper with regard to reactivity, the absence of deadlines, the exchange of participants, and the informal writing style ([16]). Comparing Rue89 and AgoraVox, the parameter of professionalism of the authors might influence the degree of subjectivity and the way it is expressed. Members of Rue89’s editorial board are professional journalists and authors submitting articles to the website are professional journalists as well, or at least so called domain specialists ([16]).

4. Texts for which this might be the case when regarding the paper version of Le Soir, such as wire copies or newsslashes, have been excluded from our corpus as explained in section 2.

5. As studies on subjectivity are often based on corpora build up of texts dealing with the same topic, we plan to compare different sub corpora in subsequent studies.

6. In the domain of corpus linguistics, the term text type was first introduced by Biber ([9], 68) who defines it by means of inner-textual linguistic characteristics, as opposed to register (previously genre), which is defined in terms of external and cultural criteria linked to the author’s purpose.

7. To determine in a more general way whether we effectively have to do with different genres or text types will need further investigation concerning other linguistic dimensions, but our pilot study already allows for the detection of tendencies.

As compared to interviews, single citations included in articles are kept and tagged in order to easily identify and exclude when wanted, as the choice of a citation reflects a personal state, not only of the source, but also of the author, and citations thus may serve our subsequent analyses of subjectivity.

As the traditional newspaper aims to treat all actual topics also in the paper version, journalists may have time pressure due to deadlines when recording the printed support as well, depending on the topic of the article.
4 Theoretical Framework and Research Method

Our quantitative analysis is based on two axes, namely (1) discourse organisation through initial position (i.e. the first preverbal zone of a given sentence) and (2) subjectivity through PSEs. While the first focuses on the evaluation of the typological differences between the media, the second is devoted to subjectivity in order to observe the variation of the author’s presence in the texts. Nevertheless, the two axes interact: Not only subjective language, but also the order of information reflects speaker stance, namely by choosing the information included, by mentioning certain aspects before others, or by linking different texts parts (phrases, sentences, paragraphs).

Before we introduce our research methods, we briefly sketch the theoretical principles our research is based on. The work in progress presented is situated in a corpus linguistic framework of discourse analysis. The present paper aims to outline first tendencies in our data. We describe and evaluate the typological differences between the three media by applying a corpus-based methodology providing a description of the global discourse organisation of our data sets and the expression of subjectivity by certain predefined cues (section 4.3).

4.1 Discourse Organisation through Initial Position

Because of the apparent incompatibility between the qualitative nature of discourse analysis and the quantitative requirements of corpus linguistics, discourse organisation is usually difficult to study by means of corpus linguistic methods ([8]). Ho-Dac ([27]) proposes a method providing a solution to this incompatibility, allowing for a data-driven approach to discourse organisation based on automatic tagging and quantitative analysis of the discourse roles of sentence-initial elements in different text positions given by the layout. The theoretically-based hypothesis is that the initial position – defined as the starting point of the message and composed of the first elements that the reader receives – has an important function in discourse organisation. The analysis of the distribution of these elements according to their text position gives an overview of the textual organisation of different text types. Therefore, two text positions are distinguished: P1 corresponding to sentences introducing a paragraph, and P2 corresponding to intraparagraphic sentences. Elements in P1 are by definition associated with a paragraph break, i.e. a visual cue of discontinuity. As a consequence, they have a greater capacity of signalling high-level discontinuities and orienting high-level segments. Because discourse organisation is complex and texts are organised according to different structuring principles, we have to consider different types of discourse segments. In this study we focus on cues that potentially signal topical continuity, rhetorical articulation, setting discontinuity, and textual discontinuity.

Topical continuity is outlined by means of co-referential grammatical subjects covering pronouns, possessive noun phrases, reiterations, and detached appositions in initial position. Several studies in cognitive linguistics showed that linguistic means available to refer to a given entity already mentioned in the text are associated with different degrees of accessibility (e.g. [41], [2], [23]). On their basis we assume that (1) co-referential expressions, especially when occurring in grammatical subject position, have an instructional meaning indicating topical continuity, and that (2) the type of this expression indicates different levels of topical continuity. For example, a first person personal pronoun in grammatical subject position indicates a strong topical continuity while reiteration may be used to reintroduce a topic or to reinforce a topical continuity when there is a discourse shift e.g. a paragraph break, a setting or a textual discontinuity cue ([46]). Another topical continuity cue is apposition, which is an attributive construction communicating supplementary information on a given sentence constituent from which it is syntactically detached. Concerning discourse organisation, and especially when occurring in initial position just before the first grammatical subject, appositions may indicate topical continuity just like to referential links ([18]), and it has been shown that the more narrative a text, the more appositions and pronouns occur in P1 ([27]).

For rhetorical articulations we only consider connectives occurring in absolute first position. When introducing a sentence or a paragraph, they may acquire a high-level discourse function in order to signal a rhetorical articulation taking place inside in the course of a given continuity (concerning topic or setting). Ho-Dac ([27]) shows that the more argumentative a text, the more connectives occur in initial position.

Setting discontinuity is outlined by means of detached setting adverbials. When occurring in sentence-initial position, setting adverbials may orient the reader by indicating the domain of applicability within which the following proposition holds (e.g. [13], [21] and [22]). In the present study, we focus on time, space, and notional adverbials, i.e. elements which set a notion that may be a domain of knowledge (in linguistics), a defined object (concerning the case of adverbials), a specific point of view (in line with Halliday), etc. The text part introduced by these adverbials is labelled discourse frame and characterized by temporal, spatial, or notional homogeneity ([17]). Ho-Dac ([27]) shows that the more descriptive a text, the more setting adverbials occur in initial position.

Concerning textual discontinuity, we focus on sequencers (linking adverbials and grammatical subjects introducing items) that serve to indicate discourse organisation attributing limits of different text parts and information sources by explicitly indicating the position of a given segment in discourse (e.g. Firstly, Secondly, Finally, Moreover, Besides, etc.).

4.2 Private State

Subjectivity generally refers to the expression of personal state, covering devices of opinion, evaluation, attitude or emotion or sentiment when generally speaking. Depending on the underlying theory and the linguistic means at focus, the phenomenon is amongst others designated as stance ([9], [11]), appraisal ([38], [51]), hedging ([35], [29]), commitment ([47]), private state ([42]) or evaluation ([5]).

Diverse means can serve to express subjectivity in texts. Usually any subjective element is linked to its emitter who can either be the writer or some other person referred to or cited in the text. In the same way, subjective elements are generally linked to a goal that the personal state relates to. In line with Thompson and Hunston ([48]), we define private state as the broad cover term for the expression of the speaker’s or writer’s attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about. That attitude may relate to certainty or obligation or desirability or any of a number of other sets of values.8

8 Connectives are more often used to link several continuous clauses inside a given sentence.

9 For further description see Bednarek ([5]).

10 While the defined phenomenon is labelled evaluation by Thompson and Hunston ([48]), we use the terms subjectivity, stance and private state as equivalents to it in the course of this article.
4.3 Potential Subjective Elements

By the second axis, we explore the use of Potential Subjective Elements (PSE) in the three data sets, accounting for occurrences of first person personal and possessive pronouns, stance adverbials, it-extrapositions, cleft sentences, and hapax, i.e. words that occur just once in a given data set.

The use of a first person pronoun is one of the most conspicuous means used to express subjectivity as it "refers to the act of individual discourse in which it is pronounced, and by which it designates the speaker" ([7]: 226) (examples (i), (iv)-(vi)). Several linguistic investigations on speaker stance – most of them including English as (at least one of) the language(s) under investigation and focusing on academic and scientific discourse – outline the importance of the choice of personal pronouns in order to express the degree of involvement in relation to the propositional content (e.g. [24], [25], [32], [33], [34], [49]). The third person singular pronoun can fulfill the communication of subjectivity when occurring as grammatical subject, too. In French, this is the case for the third person singular pronoun on used as an alternative to the first person plural pronoun nous, and taking over the speaker-inclusive meaning (examples (vii)-(viii)).

In addition to pronouns we include stance adverbials and two special constructions that are potentially related to subjectivity: it-extrapositions and cleft sentences. We are especially interested in these cues of subjective language as they offer the author the possibility to express stance in an indirect way.

Stance adverbials (like évidemment in example (ix)) can be used for reasons linked to content (e.g. when information about a topic is not sufficiently accessible), or for interpersonal reasons (e.g. when the author does not want to impose a personal point of view to the readers) ([30], [31]). Originally, stance adverbials have been defined as a means of hedging by rendering the affiliation of an object to a certain category fuzzier ([35]), while they are accredited now a more general function, including the expression of attitude, emotion and opinion. The particularity of it-extrapolations is that they express a subjective meaning while at the same time communicating a certain degree of distance between the author and the propositional content (example (ix)). As Charaudeau ([14]) points out, the use of it-extrapolation is very frequent in journalese, as these formulations seem to be less subjective, so that we speak of constructed objectivity ([16]: 504). Cleft sentences can also express an indirect judgement on the propositional content of a message ([36]). Their main function is to focus on an extracted element that is detached from the other sentence components in order to be emphasised (examples (viii), (x)). By choosing a cleft sentence structure, the author can communicate the accentuation of a given propositional content.

Lastly, we focus on single word occurrences, so called hapax legomena. Following Wiebe et al. ([52]), we assume that one word occurrences are especially interesting when exploring subjectivity in discourse. Existing studies investigating stance by a corpus-based approach are all based on English language data ([28], [15], [10], [19], [40] amongst others). Until now – to our knowledge – there are no such analyses based on French large-scale corpora.

5 Quantitative Analysis

Our quantitative analyses of the different cues presented in section 4 are based on automatic tagging which has already been used and evaluated twice ([27] and [39]).

5.1 Cues Marking

The quantitative analyses presented are based on an automatic labelling of features concerning discourse organisation on the one
hand, and subjectivity on the other hand. This labelling is based on the results of a POS tagging (TreeTagger, [45]), a syntactic parser (Syntex, [12]), and on layout information directly extracted from the TEI encoding of the corpus (section 2).

Concerning discourse organisation, the automatic marking extracts a selection of potential organisational cues occurring in initial position, distinguishing connectives occurring in absolute first position, detached elements, and grammatical subjects. These elements are automatically characterized by their POS, their function (setting vs. textual adverbials, sequencers, appositions, etc.), the semantic meaning (e.g. temporal, spatial, and notional setting adverbials), and the properties of reiteration (when an NP’s head restates a noun already mentioned in a given section). Moreover, these elements are associated with their textual position, i.e. P1 (if their host sentence introduces a paragraph) or P2 (if the host sentence is intraparagraphic)\(^{15}\). The automatic characterisation is based on a Perl script (1) delimiting the first preverbal zone for all sentences, (2) identifying all syntactic blocs composing the preverbal zone (based on Syntex results), (3) categorising each bloc by applying a set of regular expressions associated with lexical lists concerning functional and semantic features.

Detecting the PSEs, we use the POS database in order to outline hapax legomena. Concerning cues marking, we have adapted the Perl script used for discourse organisation to extract first person pronouns (je, nous, and on in subject position and mel'm', moi, nous, sel's in other positions), and possessive NPs (mon/malomes X, nos/notre X). Moreover, three other cues, namely stance adverbials and it-extrapositions as well as cleft sentences, are automatically extracted, based on the POS tagging, a lexical list for the former, and syntactic patterns for the latter.

**Table 2. Extracted Cues**

<table>
<thead>
<tr>
<th>DISCOURSE ORGANISATION</th>
<th>SETTING</th>
<th>SEQUENCE</th>
<th>CONNECT</th>
<th>APPPOS</th>
<th>COREF</th>
<th>COREFp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Setting adverbials</td>
<td>sequencers</td>
<td>connectives ( coordinations, adverbs)</td>
<td>appositions</td>
<td>proper nouns, definites, demonstratives, possessive and undetermined NPs with a syntactic head reiterating a noun already mentioned</td>
<td>pronoun and possessive NPs</td>
</tr>
</tbody>
</table>

**POTENTIAL SUBJECTIVE ELEMENTS (PSE)**

<table>
<thead>
<tr>
<th>STANCE</th>
<th>LOCpro</th>
<th>LOCposs</th>
<th>IT-ex</th>
<th>CLEFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>stance adverbials</td>
<td>1st person personal pronouns (including on)</td>
<td>NPs with 1st person possessive determiner</td>
<td>it-extrapositions</td>
<td>cleft sentences</td>
</tr>
</tbody>
</table>

### 5.2 Frequency Analysis

We firstly describe the main differences between the three data sets in terms of layout, discourse organisation, and occurrences of PSEs. Secondly, we expose occurrence frequencies of the diverse cues for each data set by using contingency tables (comparing the data sets two by two) and the log-likelihood ratio (henceforth LL\(^{16}\)) in order to measure the significant relative frequency differences between them.

The higher the LL value, the more significant the difference is between two frequency scores. In this study principally aiming at describing main tendencies, we only focus on LL corresponding to p < 0.0001 (i.e. higher than 15.13). The resulting tendencies will or will not support our hypotheses and will constitute the starting point for further detailed, quantitative and qualitative analyses.\(^{17}\) Before presenting tendencies concerning subjectivity in our corpus, the next section describe the linguistic characteristics of the three media in terms of layout and discourse organisation.

### 6 Interim Results and Tendencies

This section presents the first results and insights gained throughout the quantitative analysis. Its main concern is to expose observed general tendencies on the basis of frequency analyses of the cues for each data set and LL statistics for their comparison.

#### 6.1 Linguistic Characterisation of the Three Media

To give an overview of the general characteristics of the three data sets, we describe in the following their layout and lexical diversity. While the first has to do with discourse organisation, the second may be linked to subjectivity ([52]). Table 3 suggests different units of measurement, more or less related to layout, in order to describe their size and textual segmentation.

**Table 3. Layout Segmentation**

<table>
<thead>
<tr>
<th></th>
<th>Rue89</th>
<th>AgoraVox</th>
<th>Le Soir</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words/Article</td>
<td>564</td>
<td>751</td>
<td>467</td>
<td>582</td>
</tr>
<tr>
<td>Headsings/Article</td>
<td>50</td>
<td>68</td>
<td>42</td>
<td>53</td>
</tr>
<tr>
<td>Articles</td>
<td>3,879</td>
<td>4,368</td>
<td>5,873</td>
<td>14,200</td>
</tr>
<tr>
<td>Paragraphs/Article</td>
<td>112</td>
<td>14</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Sentences/Paragraph</td>
<td>2.7</td>
<td>3.5</td>
<td>2.5</td>
<td>2.9</td>
</tr>
</tbody>
</table>

AgoraVox is the largest data set concerning the overall number of words, paragraphs, and headings. It also contains the longest articles (on average 751 words/text) and longer paragraphs as compared to the two other media. In contrast, Le Soir shows shorter articles (on average 467 words/text) and paragraphs. As a consequence, Le Soir is the larger data set with respect to the total number of articles (5,873). Rue89, the smallest sub corpus, occupies an intermediate position.\(^{18}\) Paragraph size may play an important role in discourse organisation, allowing for simple structures in short paragraphs as compared to longer ones.\(^{19}\)

Lexical diversity is evaluated in the present by using the type/token ratio based on the idea that the more type than as compared to the number of tokens, the more varied is the vocabulary. And the closer to 1 the ratio, the more lexically diverse is the data set. Lexical diversity might be linked to theoritical presence and the expression of private state in the text as outlined in sections 3 and 4.3.

---

15 For more methodological details see Ho-Dac ([27]).
16 See [43] for details on the use of this ratio for large-scale corpus comparison.
17 Our research prospects include the investigation of more quantitative analyses on modal expressions, stance adverbials, and adjectives expressing subjectivity, as well as *verba dicendi et sentendi*. We also intend to carry out qualitative analyses on occurrences of all mentioned subjectivity cues.
18 Rue89 is the youngest media founded in 2007, explaining the comparatively smaller size.
19 Sentence length will be analysed in subsequent investigations, being another discourse organisation factor.
The type/token ratio shows that lexical diversity is more elevated in Rue89 and Le Soir, with a higher degree of diversity in Rue89. To characterise the three media in terms of their discourse organisation, we first compare them concerning the frequencies of discourse organisation cues, and second with respect to the content of P1 and P2, applying the methodology described in section 4.

Table 5 gives the LL statistics for the sentences beginning with at least one organisational cue.

<table>
<thead>
<tr>
<th>Sentences with</th>
<th>R vs. A</th>
<th>A vs. S</th>
<th>R vs. S</th>
</tr>
</thead>
<tbody>
<tr>
<td>All. org. cues</td>
<td>116.57 (R)</td>
<td>880.27 (S)</td>
<td>262.73 (S)</td>
</tr>
<tr>
<td>SETTING</td>
<td>316.12 (R)</td>
<td>427.01 (S)</td>
<td>[P2: 39.45 (S)]</td>
</tr>
<tr>
<td>SEQ</td>
<td>21.97 (A)[P2]</td>
<td>33.00 (A)[P1]</td>
<td></td>
</tr>
<tr>
<td>CONNECT</td>
<td>16.32 (A)[P1]</td>
<td>38.38 (A)[P1]</td>
<td></td>
</tr>
<tr>
<td>Topical cues</td>
<td>647.2 (S)</td>
<td>498.37 (S)</td>
<td></td>
</tr>
<tr>
<td>APPOS</td>
<td>72.4 (R)[P2]</td>
<td>1.716.73 (S)</td>
<td>832.41 (S)</td>
</tr>
<tr>
<td>COREF,t</td>
<td>91.35 (A)[P2]</td>
<td>[P1: 34.55 (S)]</td>
<td>93.73 (S)</td>
</tr>
<tr>
<td>COREF,p</td>
<td>[P1: 151.29 (A)]</td>
<td>15.64 (R)[P1]</td>
<td></td>
</tr>
</tbody>
</table>

LL statistics indicate diverse differences between the three media, and support our hypothesis concerning the typological difference between them. If we look at the first rows, Le Soir appears to be the media with the highest number of cues signalling discourse organisation. But this overuse is only effective for topical continuity cues and especially via appositions and reiterations as shown in the last three rows. Nevertheless, this overuse is not effective for all topical cues. Indeed, each topical cue is significantly associated with different media: appositions with Le Soir (and in a weaker proportion with Rue89), reiterations with AgoraVox and Le Soir, and pronouns and possessive NPs with Rue89. AgoraVox is the media with the lowest amount of organisational cues. Nevertheless, this weaker proportion of organisational cues must be qualified by looking at the detailed LL indicating that there are significantly more sequencers and connectives in AgoraVox. Concerning setting cues, Rue89 and Le Soir seem to be alike, being significantly more present in the first than in AgoraVox. If we now look at the columns, Le Soir emerges as the most specific data set in contrast to Rue89 and AgoraVox that are closer in terms of discourse organisation. Nevertheless, Rue89 and AgoraVox are not similar. They weakly differ for all different cues: (1) while AgoraVox prefers reiteration, Rue89 shows a higher amount of strong topical continuity devices (appositions and pronouns), (2) while AgoraVox shows more sequencers and connectives, Rue89 shows more setting adverbials, comparable to Le Soir.

Taking into account variations according to textual position (indicated by brackets), we find significantly more setting adverbials, sequencers, reiterations, and appositions in P1 and significantly more connectives, pronouns, and possessive NPs in P2 (Ho-Dac’s ([27]) results are in line with our insights). It is only if we focus on variations between media in each textual position that new insights appear. Setting adverbials in Le Soir are overused only in P2 when comparing Le Soir to Rue89, i.e. setting adverbials are overused in Le Soir when they are not associated with an effective structuring power ([26]). When connectives are overused in a media, it is generally in P1. In AgoraVox vs. Rue89 and Le Soir, but also in Le Soir vs. Rue89, the difference between the three data sets concerning the use of these argumentative elements is conspicuous. Pronouns and possessive NPs are overused in Rue89 when occurring in P1, underlyng global topical continuity. In contrast, it is intraparagraphic reiterations that are overused in AgoraVox.

All these observations allow for assuming that the three data sets under investigation are different. Although further analyses are needed in order to better understand the differences, we may state here that the media show more characteristics of the argumentative text type (as compared to descriptive or expository texts), considering the use of connectives in P1 (associated with argumentative text types ([27])) as compared to the use of setting adverbials in P2 (associated with descriptive text types([27])).

### 6.2 PSEs as lexico-syntactic elements

The present subsection describes results concerning the research axis on subjectivity by the use of potential subjective elements, outlined in section 4.3.

**Table 6. PSE Distribution – Number of Sentences**

<table>
<thead>
<tr>
<th>Sentences with</th>
<th>Rue89</th>
<th>AgoraVox</th>
<th>Le Soir</th>
</tr>
</thead>
<tbody>
<tr>
<td>STANCENb</td>
<td>3,229</td>
<td>6,059</td>
<td>3,201</td>
</tr>
<tr>
<td>%</td>
<td>1.65</td>
<td>2.03</td>
<td>1.28</td>
</tr>
<tr>
<td>LOCproNb</td>
<td>29,468</td>
<td>42,952</td>
<td>24,413</td>
</tr>
<tr>
<td>%</td>
<td>15.08</td>
<td>14.38</td>
<td>9.77</td>
</tr>
<tr>
<td>LOCpossNb</td>
<td>4,715</td>
<td>7,467</td>
<td>4,531</td>
</tr>
<tr>
<td>%</td>
<td>2.41</td>
<td>2.50</td>
<td>1.81</td>
</tr>
<tr>
<td>IT-exNb</td>
<td>1,219</td>
<td>2,814</td>
<td>1,103</td>
</tr>
<tr>
<td>%</td>
<td>0.62</td>
<td>0.94</td>
<td>0.44</td>
</tr>
<tr>
<td>CLEFTNb</td>
<td>4,567</td>
<td>6,256</td>
<td>4,762</td>
</tr>
<tr>
<td>%</td>
<td>2.34</td>
<td>2.09</td>
<td>1.91</td>
</tr>
<tr>
<td>total</td>
<td>195,395</td>
<td>298,636</td>
<td>249,830</td>
</tr>
</tbody>
</table>

Table 6 displays the overall occurrences of PSEs included in the present investigation for each data set. As can be seen, LOCpro constitutes the most prolific cue, with about 15% of sentences containing a personal pronoun referring to the first person in the alternative media and 9% in the traditional newspaper. All other cues occur much less frequently and divergences between traditional and alternative media are not that striking. It-extrapositions are the least frequent means for expressing subjectivity in all data sets (R = 0.62%, A = 0.94%, S = 0.44%), while first person possessive pronouns, stance adverbials and cleft sentences occupy an intermediate position with alike frequencies in the different sub corpora. It is striking that the number of sentences any of the given PSE is never higer for Le Soir than for Rue89 or AgoraVox when considering percentages.

**Table 7** represents the LL realised for the selected PSEs occurring in our corpus, comparing the data sets two by two. The first striking result is that the Le Soir data set never corresponds to the one with overuse for any of the subjectivity cues under investigation. This is in line with the comparison of percentages in table 6. Second, the divergence between alternative media on the one hand and the traditional newspaper on the other is eye-catching, especially when comparing the frequency of sentences containing a first person personal
pronoun (A vs. S LL = 2,395.78 and R vs. S LL = 2,528.77). Third, the differences comparing Rue89 and AgoraVox are much less conspicuous. Rue89 displays significantly more personal pronouns (LL LOCpro = 39.19) and cleft sentences (LL = 31.47), while AgoraVox overuses stance adverbials (LL = 90.4) and it-extrapositions (LL = 151.86). The frequency of possessive pronouns does not differ significantly between the two alternative media, in which they are respectively overused as compared to the traditional newspaper (A vs. S LL = 297.28 and R vs. S LL = 188.14). Rue89 seems to overuse cleft sentences (R vs. A LL = 31.47 and R vs. S LL = 96.66) that generally serve to point out an element by detachment, and which may as well be an indication for a more informal language style in Rue89. In contrast, the high frequency of it-extrapositions in the AgoraVox data set (R vs. A LL = 151.86 and A vs. S LL = 499) reflects an overuse of these constructions commonly associated with an impersonal expression of private state. These findings may be associated with the smaller amount of first person personal pronouns in AgoraVox as compared to Rue89. The assumption of a more informal language use in Rue89 and a more impersonal expression of subjectivity linked to it ask for further investigation.  

6.3 PSE as Hapax

Table 8 concerning subjectivity cues analyses occurrence patterns of hapax legomena.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Rue89</th>
<th>AgoraVox</th>
<th>Le Soir</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>3,131,675</td>
<td>5,092,708</td>
<td>3,836,117</td>
</tr>
<tr>
<td>Hapax Nb</td>
<td>26,849</td>
<td>38,006</td>
<td>29,777</td>
</tr>
<tr>
<td>%</td>
<td>0.86</td>
<td>0.73</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 8. Distribution and LL Statistics Concerning Hapax Legomena

<table>
<thead>
<tr>
<th>Hapax</th>
<th>R vs. A</th>
<th>A vs. S</th>
<th>R vs. S</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL statistics</td>
<td>300.17 (R)</td>
<td>25.80 (S)</td>
<td>139.16 (R)</td>
</tr>
</tbody>
</table>

As can be seen, their frequency is significantly higher in Rue89 than in the two other media (R vs. S LL = 139.16 and R vs. A LL = 300.17). AgoraVox shows the lowest type/token ratio (table 4: 0.15928) and also the lowest amount of one word occurrences as compared to the two other media (R vs. A LL = 300.17 and A vs. S LL = 25.80). But as articles published in AgoraVox have a longer mean length (table 3: A = 751.19 words/text and S = 467.27 words/text), this first tendency has to be put into perspective and controlled by further research. Because even if “people are creative when they are being opinionated” ([52]: 286), the corpus of journalism texts may show a high amount of hapax due to technical terms and specific language linked to a given subject. This might be an explanation for the low amounts of hapax and type/token ratio in Rue89, which is due to the professionalism of authors on the one hand side and the aim to report ‘differently’ from the traditional newspapers on the other, and the intermediate position of Le Soir, associated with a professional and thus probably more technical but less individual language use.

7 Conclusion

The present paper outlines the occurrences patterns of potential subjective elements in three different types of written mass media. In order to outline the expression of subjectivity, we carried out quantitative analyses by which we draw first tendencies to respond to our research questions and tested our hypotheses. The results show that the use of the different PSEs varies in the three data sets, and percentages (table 6) and raw frequencies (table 7) show that their use is less frequent in Le Soir than in the two alternative media, which is consistent with our first hypothesis. While articles in the alternative media seem to be alike, they clearly differ from the traditional newspaper. First tendencies support our second hypothesis as well: The strikingly higher use of first person personal pronouns in AgoraVox and Rue89 reflects an overt presence of the author in the these two media, as compared to Le Soir. The amount of the other PSEs under consideration is also slightly lower in the traditional newspaper. By contrast, our third hypothesis was not confirmed. Our data betoke that the two alternative media seem to prefer different PSEs, but we cannot declare an intermediate position for Rue89. Concerning the presence of the author in the text, it even seems to be more overt in Rue89, given the higher amount of personal pronouns and hapax. The high frequency of it-extrapositions in AgoraVox may indicate a subjectivity that is expressed via constructed objectivity as compared to Rue89, where overuses of cleft sentences and first person personal pronouns may be an indication for a more informal or direct expression of subjectivity. These cues will have to be investigated in subsequent research steps, including further quantitative analyses on supplementary PSE such as adjectives, verba dicendi et sentendi, or modal expressions, as well as more detailed qualitative analyses with regard to the presented PSEs and the creation of different sub corpora. Furthermore, our results support our hypothesis concerning a typological difference between the three media. The results effectively indicate differences between the three data sets, being less well-defined when comparing the two alternative media, but being conspicuous when opposing the former to the traditional newspaper.

ACKNOWLEDGEMENTS

This research was supported by grant number ARC 08-12-001 from the Belgian French Speaking Community. We thank Liesbeth De-gand and Anne Catherine Simon for proofreading and useful re-
marks as well as the three anonymous reviewers for their critical but
notwithstanding constructive comments.

REFERENCES

[5] Tonika Bien, Entities in Media Discourse: Analysis of a Newspaper
Corpus, Continuum, 2006.
[6] Francoise Benhamou, Julie Lambert, and Marc-Olivier Pudis, ‘Le jour-
nalisme en ligne: Transposition ou réinvention? entretien avec laurent
corpus analysis to describe discourse structure, volume 28 of Studies in
corpus linguistics, John Benjamins Publishing Company: Amster-
[9] Douglas Biber, Variation Across Speech and Writing, Cambridge: Cam-
[10] Douglas Biber and Edward Finegan, ‘Styles of stance in english: Lex-
ical and grammatical marking of evidentiality and affect’, Text, 9, 93–
ward Finegan, Longman Grammar of Spoken and Written English, Lon-
[12] Didier Bourigault, Un Analyseur Syntaxique Opérationnel : SYNTEX,
Ph.D. dissertation, Mémorial d'HDR en Sciences du Langage, CLLE-
[13] Wallace Chafe, Subject and Topic, chapter Givenness, Contrastiv-
ness, Definiteness, Subjects, Topics, and Point of View, 25–55, New
[14] Patrick Charaudeau, ‘Discours journalistique et positionnements énan-
[15] Maggie Charles, ‘“this mystery...”: A corpus-based study of the use of
nouns to construct stance in essays from two contrasting disciplines’,
[16] Maggie Charles, ‘Construction of stance in reporting clauses: A cross-
disciplinary study of these’, Applied Linguistics, 27(3), 492–518,
[17] Michel Charolles, ‘L'enca drement du discours; univers champs do-
maines et espaces’, Cahier de Recherche Linguistique, LandDisCo uni-
versité Nancy2, (6), (1997).
[18] Bernard Combettes, ‘Les constructions détachées comme cadres de dis-
and the Construction of Discourse, chapter Adverbial Marking of
Stance in Speech and Writing, 56–73, Oxford: Oxford University Press,
2000.
(submitted).
[22] Peter Fries, On Subject and Theme: A Discourse Functional Perspec-
tive, chapter Themes Method of Development and texts, 317–359, John
and the form of referring expressions in discourse’, Language, 69,
[24] Nigel Harwood, ‘“nowhere has anyone attempted...in this article i aim
to do just that”: A corpus-based study of self-promotional i and we in
academic writing across four disciplines’, Journal of Pragmatics, 37,
[25] Nigel Harwood, ‘we do not seem to have a theory...the theory i present
here attempts to fill this gap’: Inclusive and exclusive pronouns in aca-
[26] L.-M. Ho-Dac and M.-P. Pry-Woodley, ‘Thématologie exploratoire out-
ille pour l’étude de l’organisation du discours’, in Actes du Congres Mon-
[27] Lydia-Mai Ho-Dac, A Mosaic of Corpus Linguistics. Selected Appro-
aches., chapter An exploratory data-driven analysis for describing
[28] Susan Hunston and Geoff Thompson, Evaluation in Text, Authorial
Stance and the Construction of Discourse, Oxford: Oxford University
[32] Ken Hyland, ‘Stance and engagement: A modal of interaction in aca-
[34] Chih-Hua Kuo, ‘The use of personal pronouns: Role relationships in
scientific journal articles’, English for Specific Purposes, 18(2), 121–
[35] George Lakoff, ‘Hedges: A study in meaning criteria and the logic of
fuzzy concepts’, in Papers from the Eighth Regional Meeting, eds., Paul
Peranteau, Judith Levi, and Gloria Phares, pp. 183–228. Chicago Lin-
guistics Society (CLS 8), (1972).
[36] Knaud Lambrecht, Language Typology and Language Universals, chap-
ter Dislocation, 1050–1079, Berlin/New York: Mouton de Gruyter,
[37] Gilles Lurgin, ‘Le mélange des genres das hypertexture’, Semen, 13,
[38] James R. Martin, ‘Reading positions/positioning readers: Judgement in
C. Fabre, S. Ferrari, L.-M. Ho-Dac, A. Le Drouaile, Y. Mathet,
P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. Vergez-Couret,
L. Vieu, and A. Widlicher, ‘Annodis : une approche outillée de
l’annotation de structures discursives’, in TALN 2009, Senlis, (June
2009). ATALA, LIPN.
[40] K Precht, ‘Stance moods in spoken english: Evidentiality and affect in
[41] Ellen Prince, Radical Pragmatics, chapter Toward a Taxonomy of
[42] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik,
A Comprehensive Grammar of the English Language, London: Long-
[43] Paul Rayson and Roger Garside, Comparing corpora using frequency
profiling, 1–6, October 2000.
[45] Helmut Schmid, TreeTagger, IMS, Universität Stuttgart, Germany.
[46] Catherine Schneedecker, Noms Propres et Chaînes de référence,
[47] Geoff Thompson and Susan Hunston, Evaluation in Text, Authorial
Stance and the Construction of Discourse, chapter Evaluation: An
[48] Irena Vassileva, ‘Who am i/who are we in academic writing?’, Interna
[49] Marc Weber, Rein Vos, and R. Harald Baayen, ‘Extracting the lowest-
frequency words: Pitfalls and possibilities’, Computational Linguistics,
[50] Peter R. R. White, Appraisal Outline, www.grammaticos.com/appraisal,