doi:10.5128/ERYa8.11

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 8, 171–184

LEXICON-BASED DETECTION OF EMOTION IN DIFFERENT TYPES OF TEXTS: PRELIMINARY REMARKS

Hille Pajupuu, Krista Kerge, Rene Altrov

Abstract. Paragraphs of four genres are analysed to detect their emotional colouring, while a lexicon-based approach of linguistic analysis is weighed against reader opinion. The aim is to find out the prospects of automatic detection of emotion in any text by using a very small lexicon of about 600 frequent emotion words.*

Keywords: analysis of emotion, paragraph, reader opinion, text-linguistic analysis of emotion, word frequency

1. Introduction

The aim of the study is to make preparations for the creation of an automatic tool enabling the provision of a preliminary idea of the emotionality of any written text (including factual); that is, whether this or that text may affect the reader as positive or negative. The theoretical background relies on the cognitivist understanding of the inseparability of linguistic and other experience (see e.g. Langacker 2000 and other articles in Barlow, Kemmer 2000 on usage-based models of language) and on views of the recent 30 years on the importance of text reception. According to the latter we define text as a unit of communication to be interpreted by the receiver – relying, of course, on the norms of genre and usage (see e.g. de Beaugrande, Dressler 1981/2002, Kecskes 2008). Therefore the study is centred on reader opinion, whether or not the reader has any expert knowledge of what the emotive effect might be due to (certain linguistic expressions, personal or general background of knowledge and cognition, a fact, argument, emotion or opinion contained in the text).

Determining the emotionality of texts with the help of a dictionary, we rely on the principles of lexicon-based detection of emotion as used in sentiment analysis (opinion mining) for various systems, which means that the text words carrying an emotional colouring are labelled accordingly using dictionaries or word lists where the words have been provided with emotion tags either automatically or manually

* We are thankful to the anonymous referees for their vital remarks and the Projects EKT1, SF0050023s09 and ETF8605.

(see Taboada et al. 2011). In different dictionaries word emotion may be described in layers; that is, using different dimensions of emotions, one of which is polarity, defining words as positive or negative. In systems focused on text polarity, texts are searched for positive and negative words, tagging them accordingly, while the final score of text emotionality is received by adding up the respective polarities (Missen et al. 2009).

Our study aims to test the prospects of determining the positive or negative colouring of any written text by using a very small lexicon containing frequent words with a clearly sensed polarity. Unlike in sentiment analysis that seeks to pinpoint the personal attitude of the text writer, our aim is to find out to what extent the text polarity established by the lexicon-based method might coincide with the reader's impression of the emotionality of the text.

One of the problems is the optimal length of the text subjected to an emotionality measure. As different units of a full text, such as sentences or paragraphs, may carry conflicting polarities, a full-text score need not be a too precise measure of emotionality (Pang, Lee 2008). Proceeding sentence by sentence, however, need not yield any better results, because in some cases even a human reader may find it difficult to decide over the polarity of this or that sentence (e.g. *I know he is not a good boy but he is not that bad too*; Missen et al. 2009).

In the present study the basic text unit subjected to emotionality measurement is an orthographic paragraph as a computer cannot be expected to identify substantial paragraphs. We assume that orthographic paragraph is an important functional and meaningful unit of text, one within which conflicting emotions seldom occur. However, interpretation of the results requires qualitative attention to be paid to the focus of the paragraph, making sure that the emotion of the final sentences coincides with that of the whole paragraph, because presumably the emotion perceived at the end of the paragraph has a vital effect on the reader's perception of its general polarity.

2. Procedure, material and method

The research procedure consists of three stages. In stage one we check if orthographic paragraph is an appropriate text unit for emotion detection. So we searched the Estonian Emotional Speech Corpus for such journalistic passages that had been included in the corpus only after being subjected to a reading test and getting a unanimous (either positive or negative) polarity score from all readers participating (see Altrov, Pajupuu 2008). In addition each sentence of the paragraphs was isolated from the context and was subjected to another reading test where the testers were asked to label the emotional colouring of the sentence choosing from *positive*, *negative*, and *neutral*. The results were analysed to learn the emotions dominating the paragraphs chosen and the possible change of emotion within the paragraphs.

The following stages involved analysis – using different methods – of paragraphs representing the following four different genres: literary diary (Ristikivi 2008); economy news; editorials of a daily; and the weekly horoscope of *Arter*, a weekly attachment to the *Postimees* daily. The preliminary genre selection is motivated by the conviction that one should start with studying most different kinds of

media texts, because although many text studies tend to take journalistic language as something homogeneous, in our case genre differences may affect the results. In addition it could be useful to take, as a parallel test case, a text sample from a different domain of language use where emotional polarity is supposedly manifested. As diaries looked promising in this respect we turned to Varrak publishers who provided research access to digitised parts of Ristikivi's diary, proportionally representative of each period involved.

The aim of stage two was to ascertain the emotionality of the paragraphs by using the lexicon-based method; that is, to decide for each paragraph whether its general emotionality is positive, negative, or ambivalent (an equal number of positive and negative words) or neutral (no positive or negative words). For the material analysed see Table 1.

Table 1. Material

Material	Passages from the Emotional Speech Corpus		Passages of different genres			
	Positive	Negative	Literary diary	Economy news	Horoscope	Editorials
Number of passages	18	18	17	15	12	14
Number of sentences	143	140	59	53	41	71
Number of tokens	1,189	1,322	651	812	547	1,193

In the lexicon-based detection of paragraph emotion we used *The Basic Estonian Dictionary* (Kallas, Tuulik 2011) containing the 3,015 most frequently used Estonian words; 639 of those have been labelled as having an emotional meaning, falling into 317 positive and 322 negative words. In tagging the words, not only semantics (e.g. *rõõm*^{pos} 'joy', *kuritegu*^{neg} 'crime', *ilus*^{pos} 'beautiful') but also the cultural aspect has been considered, because a word evoking positive or negative feelings in one language need not do the same in another (e.g. in Estonian words such as *kodumaine*^{pos} 'homemade', *leib*^{pos} 'bread', *sõltumatu*^{pos} 'independent', *tagasihoidlik*^{pos} 'modest', *hapu*^{neg} 'sour', *suitsetamine*^{neg} 'smoking' have an emotional colouring; see also Balahur, Montoyo 2008; on the principles of labelling emotion words in the Estonian Base Word Dictionary, see Vainik forthcoming). Our decision to use frequent emotion words only is based on the knowledge that ~60% of the words of any Estonian written text belong to the 3,000 most frequent Estonian words (Pajupuu et al. 2010). In addition, an analysis of the text corpora used by the compilers of a frequency dictionary also show that the 3,000 most frequent words cover about 64% of texts (Kaalep, Muischnek 2002). Hence we assume that every emotional paragraph would probably contain a couple of frequent emotion words.

Accordingly, each word of a paragraph was compared with the entry words of the afore-mentioned base word dictionary and tagged as positive or negative according to the emotion marker in the dictionary. The rest of the words were regarded as neutral.

As negation changes word polarity the following rules were applied (see Li et al. 2010, Pajupuu et al. forthcoming):

- by negation a neutral word (without an emotion tag) is turned negative; for example *hindama* 'to appreciate' (0), *ei hinda* 'do/does not appreciate' (–);

- by negation the following positive word is turned negative; for example, *rõõmustama* 'to be happy (about)' (+), *ei rõõmusta* 'is/are not happy' (–);
- by negation the following negative word is turned positive; for example, *valetama* 'to lie' (–), *ei valeta* 'do/does not lie' (+).

Next the positive and the negative words of a paragraph were added up. The emotion score of the paragraph depended on whether it was dominated by positive or negative words. If, for example, there was one positive and four negative words, the paragraph's emotion score was –3; that is, negative. If the number of positive and negative words was equal, the paragraph was classified as ambivalent. If there were no positive or negative words, the paragraph was regarded as neutral.

(1) Lexicon-based analysis of a weekly (*Arter*) horoscope (see full translation under example (2)):
Isegi kui üritad endast parima 'the best' (+) anda, ei pruugi 'need not' (–) kõik su soovide kohaselt laabuda. Nii mõnigi asi ei sõltu 'does not depend' (–) ainult sinust, need, kelle käes on võim ja otsustusõigus, ei pruugi 'need not' (–) olla sinuga ühel meelel. Teiste inimeste tegude eest ei saa 'cannot' (–) sina vastutust enda õlgadele võtta, ometi võib nende tegevus sulle nüüd üksjagu mõju avaldada. (Polarity –3, negative)

The aim of stage three was to test the correctness of the results of the economical, lexicon-based method described earlier and the prospects of its use in automatic analysis requiring the most optimal solutions. Here the results of lexicon-based analysis were compared with those of linguistic expert analysis, while the comparison was conducted by a native Estonian specialist of linguistic text analysis. There follows a comparison of the two methods based on reader opinion (for details see below).

The qualitative linguistic text analysis (expert analysis) was carried out in order to find out whether identification of the 'human' semantic orientation (see Taboada et al. 2011) by relying on all levels of language use revealed in the paragraphs in question could possibly yield results that are considerably better than the results of using a rather minimalistic frequency-based emotion dictionary. Thus a native specialist in linguistic text analysis read each paragraph twice, with a reasonable time interval, marking the units (words, expressions, morphological forms, phrases, or sentences) bearing polarity; that is, having a positive or negative colouring in that particular context, and finally checking whether there was sufficient coincidence between the two readings.[1] The factors that were considered also included the intensity of the emotional colouring, which may be carried by an intensifier (cf., e.g. halb 'bad' and väga halb 'very bad'), a higher degree of comparison (cf. ilus 'beautiful' and ilusam 'more beautiful'), reference to an authoritative opinion (e.g. paljude arvates 'according to general opinion'), coordination of polarity-bearing linguistic units (e.g. tüütu ja väsitav 'tedious and tiresome'), a certain type of a clause (e.g. isegi kui üritad endast parima anda 'even if you are doing your best' amplifying a negative statement), or markers of subjective modality (e.g. kindlasti 'certainly', framing the main clause mind ajas vihale 'I hated'). As a result the emotion-bearing unit may score +1 (positive emotion) or even +3 (positive emotion twice amplified). The results were added up for each paragraph.

[1] The whole procedure has been described in detail by Krista Kerge, who is the expert for the present study, in her paper "Ristikivi päevaraamatu tundetoon" ('The emotional tonality of Ristikivi's diary') delivered at a joint conference on text issues (see Kerge 2010), but the topic is beyond the limits and relevance of the present article. For one tester, the coincidence of two tests separated by a time interval should be at least 80%. In our case the double procedure has yielded sufficient coincidence. See for example resources related to content analysis and text analysis: http://www.content-analysis.de/ (accessed 24.09.2011).

For a better comparison of the paragraphs the emotion score was calculated on a different basis than with the lexicon, namely, as a rate of summed polarity to length of paragraph (the percentage of the total emotionality reading of the number of textual words in the paragraph was rounded to the nearest whole number). Such calculation – however conditional – was expected to give a better idea of the extent of the influence of the polarity-bearing words or expressions. In the observational results of the expert a score of qualitative analysis under 10 meant, conditionally, a weak polarity, 10–29 signalled of a medium one, while 30 and over meant high polarity.

For example, the Estonian version of example 2 below was a paragraph of 53 text words with a total of one positive marker and five negative ones, thus scoring –8; that is, 'slightly negative':

[(+1–5)/53] × 100 = –7.547; emotion score –8.

Note that, unlike in the lexicon-based analysis, the expert opinion is affected by two different contexts: first, the personal cognitive context of the expert, and second, that of the preceding and following text (for more on context see e.g. Kecskes 2008). In some cases the expert found the context ambivalent, resulting in alternative scores (0/–1, or neutral/slightly positive), as seen in the tables 2–5 below, the choice depending on interpretation.

The results of the two methods can be compared on example (1) (above) and example (2) (below).

(2) Linguistic text analysis of a weekly (*Arter*) horoscope:
Even if you try your best, you need not succeed in everything (–2). There are things that just do not depend on you (–1); those having power and authority need not agree with you (–1). Although you cannot take responsibility for other people's actions (+1) they may affect you quite considerably now (–1). (Polarity –4, sentiment score –8, slightly negative)

Next the results of the lexicon-based and expert analyses are compared to reader opinion, which provides, of course, the most reliable reference, but is hardly accessible to automatic analysis.

To extract reader opinion, at least five native Estonian readers were asked to read the paragraphs independently and decide by intuition whether a paragraph was positive, negative or neutral. The final result for each paragraph was gained by using the dominant opinion (Pennebaker et al. 1997). In the case of three alternative options (positive, negative, neutral) the dominant opinion was the one expressed by the readers more times than the other two together. For example, if four out of six readers assessed a passage as positive, one as negative and one as neutral, the dominant opinion defined the passage as positive. If no opinion dominated over the sum of the rest, the paragraph was regarded as emotionally ambivalent. (For example, there was no dominating opinion if three of the six readers thought that a certain passage was positive, while two marked it as negative and one as neutral.)

Both the lexicon-based and linguistic expert analyses were validated against reader opinion as it is, after all, up to the reader to interpret the meaning of a written text. If the assessments gained by different methods diverged, the material was returned to qualitative analysis, trying to guess the reasons (possibly applicable in later research). In the case of ambivalent reader opinion it becomes irrelevant

whether or to what extent the lexicon-based or expert results coincide with it. Emotionality assessment by means of any other method can be considered correct only if its results coincide with the opinion of part of the readers of the passage (for details see part 4). In this case all that needs explaining at the interpretation of the results is the mutual coincidence or divergence of the lexicon-based and linguistic text analyses.

3. Stage one: Emotion in an orthographic paragraph

The aim of this stage was to ascertain whether an orthographic paragraph was emotionally consistent enough to qualify as our research object.

According to an analysis of the paragraphs drawn from the Emotion Corpus, most of the sentences bear the same emotion as the paragraph in general. If a corpus paragraph had been marked as negative, most of its sentences were marked as negative in a context-free test (see Figure 1). If a corpus paragraph had been marked as positive, most of its sentences were marked as positive in a context-free test as well (see Figure 2). For a summary survey of all of the paragraphs analysed see Figure 3.

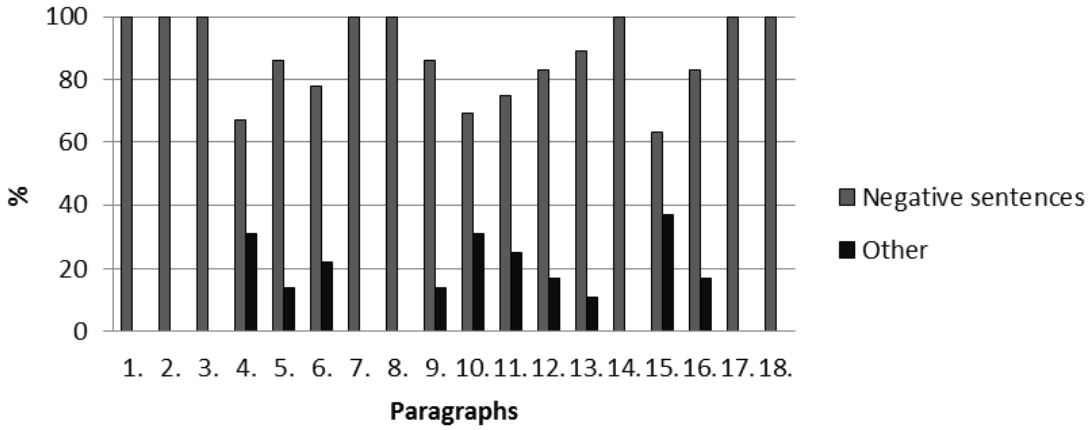

Figure 1. Negative vs. other sentences in paragraphs classified as negative

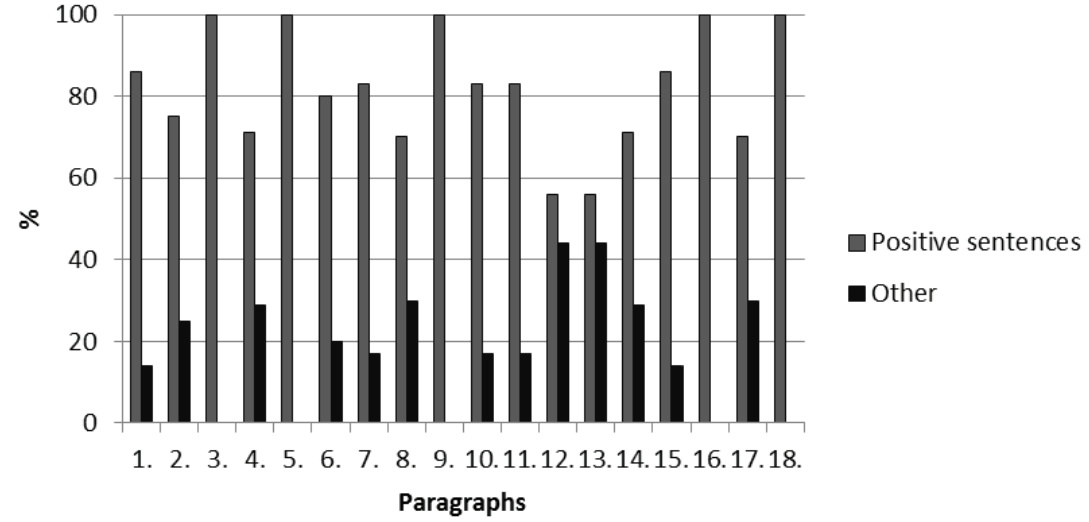

Figure 2. Positive vs. other sentences in paragraphs classified as positive

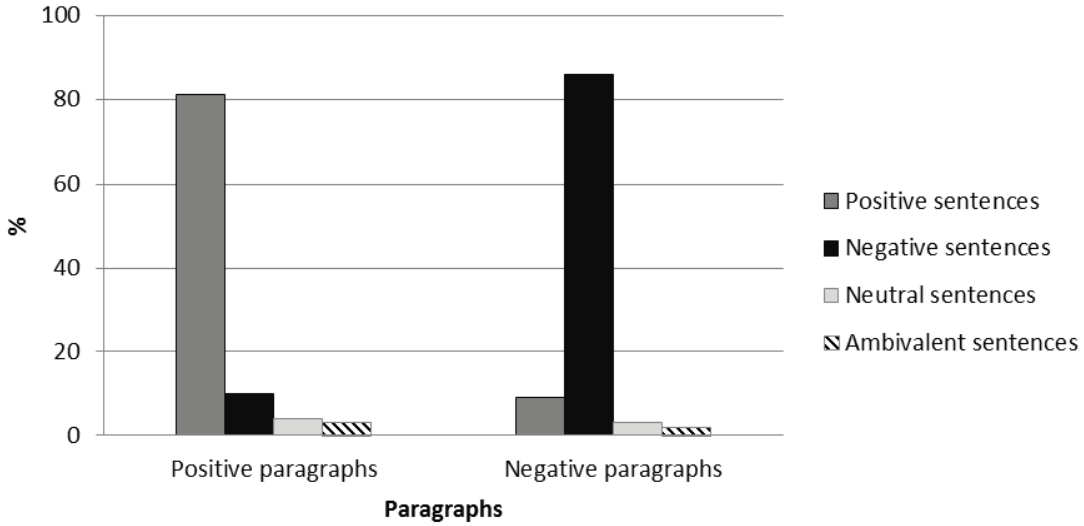

Figure 3. Assessment of sentence emotion in positive vs. negative paragraphs

The results enable the conclusion that, as far as emotional colouring is concerned, a paragraph is united enough to be treated as a unit in our analysis.

4. Stages two and three: Results of the emotion analysis of paragraphs of different genres

The results of the emotionality analysis of paragraphs of different genres are presented in Tables 2, 3, 4 and 5 as compared with reader opinion. In this analysis reader opinion is regarded as the reliable (correct) reference value (see the arguments earlier). Thus the results obtained by the lexicon-based method as well as those of linguistic text analysis were evaluated against reader opinion as correct. If the reader opinion has been marked in the table as ambivalent, it means that no reader opinion really dominates over the others, as different readers have assessed the emotionality of the paragraph differently; either: 1) some negative, some positive, some neutral, while none of the three opinions surpasses the sum of the rest; or 2) the number of negative opinions equals that of the positive; or 3) the number of negative opinions equals that of the neutral; or 4) the number of positive opinions equals that of the neutral.

In the case of ambivalent reader opinion the results obtained by either of the other two methods is regarded as correct if the ambivalence is due to polar assessments (points 1 and 2 above) because, in this case, the assessment depends on the reader, not on the lexical or some other linguistic content of the text. If the ambivalent opinion is the result of a combination of neutral assessments and those of just one pole (points 3 and 4), the results obtained by either of the other two methods is regarded as incorrect if those are opposite to the reader's opinion (i.e. if the reader's opinion is ambivalent because the text has scored an equal number of neutral and positive (resp. negative) assessments, but another method results in a 'negative' (resp. 'positive') total score, the latter result is considered incorrect as it does not match the reader opinion even partly).

The lexicon-based emotion of the literary diary (*Ristikivi päevaraamat*) coincided with reader opinion in 76.5% of the cases analysed, while the coincidence of the opinions of the text expert and non-specialist readers was 88.2%. The mutual coincidence of the lexicon-based and linguistic text analyses was 76.5% (see Table 2). The divergent opinions depend on a wider context (such as recurrence of a neutrally worded fact of everyday life in diary entries of different days (e.g. *Raadioreparaatorit täna ei tulnud* 'The radio repairman did not come today'*; Raadioreparaatorit ei tulnud ka täna* 'The radio repairman did not come today either'), which need not be available to a paragraph tester.

The grey background in the tables set off the results differing from reader opinion, while the white script against black is used for those reader opinions that coincide with the emotion determined by means of the two other tests.

Table 2. Emotion analysis of paragraphs from a literary diary

Paragraph	Lexicon-based approach		Reading test	Linguistic text analysis	
	Lexicon-based assessment	Score	Reader opinion	Linguist's assessment	Score
1.	ambivalent	0	positive	slightly positive	+5
2.	positive	+3	ambivalent	slightly positive	+6
3.	negative	–3	negative	negative	–17
4.	negative	–2	negative	highly negative	–32
5.	negative	–1	negative	slightly negative	–2
6.	negative	–2	negative	negative	–23
7.	positive	+1	positive	highly positive	+30
8.	negative	–4	negative	slightly negative	–3
9.	negative	–1	neutral	negative	–10
10.	negative	–1	negative	negative	–16
11.	ambivalent	0	negative	negative	–22
12.	negative	–1	negative	slightly negative	–7
13.	neutral	0	negative	slightly negative	–9
14.	negative	–3	negative	negative	–10/11
15.	negative	–2	negative	negative	–14
16.	negative	–1	negative	slightly positive / neutral	+1/0
17.	positive	+1	positive	slightly positive	+6

For economic news, the lexicon-based assessment coincided with reader opinion in 73.3% of the cases, while coincidence between the assessment of a text expert and reader opinion was 80.0%. The results of the lexicon-based and linguistic text analyses coincided in 46.7% of the cases (see Table 3). Besides one ambivalent paragraph assessment, differences appeared in the case of some paragraphs where critical analysis (see e.g. Fairclough 2010: 121–154, 164 ff.) revealed reader manipulation or irony. Although the 46.7% coincidence between the results of the last two methods

is not very high, the 73.3% coincidence between the lexicon-based detection (which is the method being tested) and reader opinion as the most important touchstone seems good enough, all the more so because a specific kind of news could easily use some words that do not belong to a basic dictionary.

Table 3. Emotion analysis of economic news

Paragraph	Lexicon-based method		Reading test	Linguistic text analysis	
	Lexicon-based assessment	Score	Reader opinion	Linguist's expert opinion	Score
1.	positive	+1	positive	positive	+14
2.	ambivalent	0	positive	slightly positive	+4/+5
3.	neutral	0	positive	positive	+13
4.	positive	+1	negative	negative	–13
5.	neutral	0	neutral	slightly negative	–8
6.	positive	+1	positive	positive	+11
7.	positive	+5	negative	neutral	0
8.	negative	–1	ambivalent	slightly positive	+4
9.	negative	–6	negative	negative	–10
10.	neutral	0	neutral	neutral / slightly positive	0/+5
11.	positive	+6	positive	slightly negative	–9
12.	positive	+8	positive	positive	+11
13.	negative	–2	negative	negative	–22
14.	positive	+1	ambivalent	slightly negative	–9
15.	positive	+2	positive	positive	+14

For the horoscope, the lexicon-based result coincided with reader opinion in 75.0% of the cases, while the expert–reader agreement was 58.3%. The results of the lexicon-based method and linguistic text analysis coincided in 50.0% of the cases (see Table 4).

Our material suggests that, in some genres, the objective analysis of an expert linguist may yield results differing from reader opinion. Horoscope recommendations, for example, are neither negative nor positive from a linguist's point of view, but they nevertheless rely on the partly negative premise that the recommended action is not typical of the reader (e.g. the instruction to *plan systematically* assumes that this is not something the reader usually does). In some cases the expert's opinion was biased by the fact that the general emotional colouring of the paragraph was neutralised by the strong opposite polarity of the focus at the end of the paragraph. Possibly, the reader of a relatively manipulative horoscope text will rather be guided just by certain keywords whose polarity can also be found easily by using the lexicon-based method.

Table 4. Emotion analysis of a weekly horoscope

Paragraph	Lexicon-based method		Reading test	Linguistic text analysis	
	Lexicon-based assessment	Score	Reader opinion	Linguistic assessment	Score
1.	positive	+4	positive	slightly negative	–2
2.	positive	+3	ambivalent	slightly negative	–5
3.	negative	–1	negative	neutral	0
4.	negative	–3	negative	slightly negative	–8
5.	negative	–2	positive	slightly negative	–4
6.	positive	+6	positive	highly positive	+35
7.	negative	–1	neutral	negative	–12
8.	ambivalent	0	positive	slightly negative	–4
9.	negative	–4	negative	negative	–10
10.	positive	+2	ambivalent	slightly negative	–2
11.	positive	+1	ambivalent	negative	–24
12.	negative	–3	ambivalent	slightly negative	–8

For an editorial, both the lexicon-based and expert assessment coincided with reader opinion in 71.4% of the cases. The mutual coincidence between the results of the lexicon-based method and linguistic text analysis was 85.7% (see Table 5).

Table 5. Emotion analysis of paragraphs of a daily's editorial

Paragraph	Lexicon-based method		Reading test	Linguistic text analysis	
	Lexicon-based assessment	Score	Reader opinion	Linguistic assessment	Score
1.	negative	–5	negative	slightly negative	–8
2.	negative	–7	negative	negative	–10
3.	positive	+4	positive	positive	+10
4.	negative	–4	negative	negative	–13
5.	negative	–6	negative	negative	–14
6.	neutral	0	negative	slightly positive	+2
7.	negative	–4	negative	slightly negative	–4
8.	neutral	0	positive	neutral / slightly negative	0/–1
9.	negative	–4	positive	neutral	0
10.	negative	–4	negative	negative	–10
11.	negative	–1	negative	slightly negative	–3
12.	positive	+1	negative	slightly positive	+2
13.	negative	–6	negative	slightly negative	–7
14.	negative	–11	negative	negative	–15

According to qualitative analysis the few divergent results for the editorial can be accounted for by irony or double negation (*ei ütle, et ei ole* 'will not say it is not') noticed by in-depth analysis only.

5. Discussion and conclusions

This study proves that automatic detection of the emotion of Estonian paragraphs by using a lexicon of about 600 frequent emotion words gives rather good results. For all four genres analysed the coincidence of the lexicon-based method and reader opinion was over 70%. This is comparable with other lexicon-based detectors of emotion where, depending on the lexicon and text type, the rate of correct detection can be anything between 53% and 80% (Taboada et al. 2011).

The lexicons used in emotion detection can be very different in size, with most of them containing anything between 5,000 and 70,000 words. Analyses have shown that bigger lexicons do not necessarily mean better detection rates – rather the opposite. Hitherto the best results of lexicon-based detection of emotion (78% on average) have been reported for the SO-CAL system (Semantic Orientation CALculator; see Taboada 2011: 270 ff.), which uses part- of-speech lists considering both frequency and domain and combines them into a lexicon with a total of under 5,000 entries (Brooke et al. 2009, Taboada et al. 2011). Our success with a considerably smaller lexicon can be accounted for by the fact that most of the frequent emotion words are monovalent, so that their emotional connotation is seldom changed by context (e.g. *koostöö* 'cooperation' – which is no. 349 in the frequency dictionary of Kaalep, Muischnek (2002) – is invariably positive, whereas the relatively rare word *vähenõudlik* 'modest; frugal; indulgent' can be positive or negative depending on the context).

Comparison with linguistic text analysis showed that, although in-depth analysis may be more precise, the difference is hardly big enough to discard the simple lexicon-based method and start automatising multi-level linguistic analysis. Moreover, we have observed that, for some genres, expert analysis may diverge from reader opinion even more than lexicon-based detection of emotion. Linguistic analysis of text emotion may nevertheless yield interesting results in comparative studies of genres, texts and authors. The present study indicates that differences in emotion perception are relatively big for the consciously manipulative horoscope genre (many ambivalent reader assessments and under 50% coincidence in computer vs. human assessments), whereas computer–human agreement was strongest for the diary.

The lexicon-based method should be automatised and subsequently tested on considerably more material of more genres. In addition it should be considered whether just one summary score is a sufficient parameter to characterise text emotion. According to some studies the emotion of a text generally seeming positive or negative can be reversed by the emotion of its first or last sentence, while in some other cases the summary assessments of sentence emotion diverge from that given to the whole text for other reasons (see Bestgen 1994: 11 ff., Polanyi, Zaenen 2006). The material used in our experiment turned out to contain paragraphs where the focus emotion contrasted with the emotional colouring of the rest of

the text. Therefore it might be useful to show the emotional dynamics of the whole paragraph from beginning to end.

References

Altrov, Rene; Pajupuu, Hille 2008. The Estonian emotional speech corpus: Release 1. – František Čermak, Rūta Marcinkevičienė, Erika Rimkutė, Jolanta Zabarskaitė (Eds.). Proceedings of the Third Baltic Conference on Human Language Technologies: The Third Baltic Conference on Human Language Technologies. Vytauto Didžiojo Universitetas, Lietuviu Kalbos Institutas, Vilnius, 9–15.

Balahur, Alexandra; Montoyo, Andres 2008. Applying a culture dependent emotion triggers database for text valence and emotion classification. – Procesamiento del Lenguaje Natural, Revista 40, 107–114.

Barlow, Michael; Kemmer, Suzanne (Eds.) 2000. Usage-based Models of Language. Stanford, California: CSLI Publications, Centre for the Study of Language and Information.

Bestgen, Yves 1994. Can emotional valence in stories be determined from words? – Cognition & Emotion, 7 (1), 21–36.

Brooke, Julian; Tofiloski, Milan; Taboada, Maite 2009. Cross-linguistic sentiment analysis: From English to Spanish. – International Conference RANLP 2009, 50–54.

de Beaugrande, Robert-Alain; Dressler, Wolfgang Ulrich 1981/2002. Introduction to Text Linguistics. London: Longman, 1981. Digitally reformatted 2002. http://www.beaugrande.com/introduction_to_text_linguistics.htm (2.01.2012).

Fairclough, Norman 2010. Analysing Discourse. Textual Analysis for Social Research. Oxon, New York: Routledge.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik. Tartu: TÜ kirjastus.

Kallas, Jelena; Tuulik, Maria 2011. Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 59–75. http://dx.doi.org/10.5128/ERYa7.04

Kecskes, Istvan 2008. Dueling contexts: A dynamic model of meaning. – Journal of Pragmatics, 40 (3), 385–406. http://dx.doi.org/10.1016/j.pragma.2007.12.004

Kerge, Krista 2010. Ristikivi päevaraamatu tundetoon. – Ms. Ettekanne TÜ ja TLÜ tekstipäeval 2010, 9.12.2010. Tallinn: Tallinna Ülikool.

Langacker, Ronald W. 2000. A dynamic usage-based model. – Michael Barlow, Suzanne Kemmer (Eds.). Usage-Based Models of Language. Stanford, California: CSLI Publications, Centre for the Study of Language and Information, 1–63.

Li, Shoushan; Lee, Sophia Y. M.; Chen, Ying; Huang, Chu-Ren; Zhou, Guodong 2010. Sentiment classification and polarity shifting. – Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing: Coling 2010 Organizing Committee, 635–643.

Missen, Malik Muhammad Saad; Boughanem, Mohand; Cabanac, Guillaume 2009. Challenges for sentence level opinion detection in blogs. – Eighth IEEE/ACIS International Conference on Computer and Information Science, 347–351. http://dx.doi.org/10.1109/ICIS.2009.190

Pajupuu, Hille; Kerge, Krista; Altrov, Rene (forthcoming). Detecting emotional valence of text by using a small dictionary. – Empiricism and Analytical Tools for Applied Linguistics in the 21st Century – Empirismo y herramientas analiticas para la Linguistica del Siglo XXI. Salamanca: Universidad de Salamanca.

Pajupuu, Hille; Kerge, Krista; Meister, Lya; Asu, Eva Liina; Alp, Pilvi 2010. Natural speaking and how to assess it. – Trames: Journal of the Humanities and Social Sciences, 59 (2), 120–40. http://dx.doi.org/10.3176/tr.2010.2.02

Pang, Bo; Lee, Lilian 2008. Opinion mining and sentiment analysis. – Foundations and Trends® in Information Retrieval 1–2, 1–135. http://dx.doi.org/10.1561/1500000001
Pennebaker, James W.; Mayne, Tracy J.; Francis, Martha E. 1997. Linguistic predictors of adaptive bereavement. – Journal of Personality and Social Psychology, 72 (4), 863–871. http://dx.doi.org/10.1037/0022-3514.72.4.863
Polanyi, Livia; Zaenen, Annie 2006. Contextual valence shifters. – James G. Shanahan, Yan Qu, Janyce Wiebe (Eds.). Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, 20. Dordrecht: Springer, 1–10.
Ristikivi, Karl 2008. Päevaraamat 1957–1969. Tallinn: Varrak.
Taboada, Maite; Brooke, Julian; Tofiloski, Milan; Voll, Kimberly; Stede, Manfred 2011. Lexicon-based methods for sentiment analysis. – Computational Linguistics, 37 (2), 267–307. http://dx.doi.org/10.1162/COLI_a_00049
Vainik, Ene 2012. Kuidas määrata eesti keele sõnavara tundetooni? – Eesti Rakenduslingvistika Ühingu aastaraamat, 8, 257–274. http://dx.doi.org/10.5128/ERYa8.17

Hille Pajupuu (Institute of the Estonian Language). Fields of research: speech acoustics, speech communication, national stereotypes, language testing and assessment, emotional speech and text.
hille.pajupuu@eki.ee

Krista Kerge (Tallinn University). Fields of research: linguistic varieties; discourse, genre, and text analysis; applied linguistics (e.g. L1 and L2 teaching and testing, legal and administrative communication and linguistic norms, speech synthesis, etc.).
krista.kerge@tlu.ee

Rene Altrov (Institute of the Estonian Language). Fields of research: intercultural communication, speech perception, emotional speech and text.
rene.altrov@eki.ee

ERI TÜÜPI TEKSTIDE EMOTSIONAALSUSE TUVASTAMINE LEKSIKONIPÕHISEL MEETODIL: ESMASED KATSETUSED

Hille Pajupuu[1], **Krista Kerge**[2], **Rene Altrov**[1]
[1]Eesti Keele Insituut, [2]Tallinna Ülikool

Artiklis analüüsitakse leksikonipõhisel, teksti lingvistilise analüüsi ja domineeriva lugejahinnangu meetodil neljast žanrist tekstilõikude emotsionaalsust: kirjanduslik päevaraamat (Ristikivi 2008), majandusuudised, nädalalehe Arter horoskoop, päevalehe juhtkirjad (vt tabel 1). Uurimuse eesmärk on välja selgitada, kui perspektiivikaks võiks osutuda millise tahes teksti emotsionaalsuse automaatne määramine väga väikse leksikoni abil, mis sisaldab vaid u. 600 sagedast emotsioonisõna.

Esimeses etapis tehti kindlaks, kas tekstilõik sobib uurimise objektiks. Sel eesmärgil lasti eesti emotsionaalse kõne korpuse varem määratud emotsiooni kandvate (s.o positiivsete ja negatiivsete) lõikude lausetes kontekstivabalt määrata emotsiooni poolus (positiivne, negatiivne või neutraalne). Osutus, et lõik on sobiv uurimisüksus: lausete emotsioonimäärang kattub suuresti selle lõigu emotsiooniga, millesse laused kuuluvad (vt jooniseid 1–3).

Teises etapis määrati eelnimetatud žanriti tekstilõikude emotsionaalsus eesti keele põhisõnavara sõnastiku abil (Kallas, Tuulik 2011), mille 3015 kõige sagedamast sõnast on 317 märgendatud positiivset ja 322 negatiivset emotsionaalset tähendust kandvaks. Tekstisõna sai pluss- või miinuspunkti kattumisel sõnastiku vastava emotsioonisõnaga, kuid nii, et eitus muutis neutraalse sõna negatiivseks, polaarse sõna aga vastandmärgiliseks. Lõigu emotsionaalsusskoori arvutamisel summeeriti pluss- ja miinusmärgiga sõnad.

Kolmandas etapis kasutati kaht meetodit. Asjatundja teostatud lingvistiline tekstianalüüs arvestas erinevalt sõnastikust kõigi tasandite keelendeid, nende intensiivsust, lingvistilist ja laiemat konteksti. Emotsionaalsusskoor arvutati taas pluss- ja miinuspunkte summeerides, kuid ühtlasi võeti arvesse lõigu pikkust, mis andis võimaluse rääkida lõigu nõrgast, keskmisest ja tugevast emotsioonist. Lugejahinnangu andis iga žanri tekstilõikudele vähemalt viis lugejat. Seejärel võrreldi nii sõnastikupõhise meetodi kui ka lingvistilise tekstianalüüsi tulemuste kattuvust lugejahinnanguga – viimane on kõige usaldusväärsem ehk õige, sest just lugeja annab tekstile tähenduse. Muudel meetoditel saadud hinnangud loeti õigeks siis, kui lugejahinnang oli osutunud vasturääkivaks (st ükski hinnang ei domineerinud teiste üle) ja muul meetodil saadud hinnang kattus osa lugejate hinnanguga. Võrdluse tulemused kajastuvad ülal mainitud žanrite järjestuses tabelites 2–5.

Selgus, et väike emotsioonisõnastik on tekstilõikude emotsiooni määramisel tõhus vahend: sõnastiku abil saadud määrang kattus lugejahinnanguga sõltuvalt žanrist 71,4–76,5 protsenti, mis pole halvem teiste uurijate tulemustest (Taboada jt 2011 järgi 53–80%). Lingvistilise asjatundjaanalüüsi erinevus sõnastikupõhisest analüüsist polnud nii märkimisväärne, et tasuks keeruka ja mitmetasandilise analüüsi automatiseerimisele mõelda.

Võtmesõnad: emotsioonianalüüs, lõik, lugejaarvamus, teksti lingvistiline analüüs, sõnasagedus