View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by BOP Serials

Towards a corpus-based analysis of evaluative scales associated with *even*

Volker Gast and Christoph Rzymski (Jena)

Abstract

Scalar focus operators like *even*, *only*, etc. interact with scales, i. e., ordered sets of alternatives that are referenced by focus structure. The scaling dimensions interacting with focus operators have been argued to be semantic (e. g. entailment relations, probability) in earlier work, but it has been shown that purely semantic analyses are too restrictive, and that the specific scale that a given operator interacts with is often pragmatic, in the sense of being a function of the context. If that is true, the question arises what exactly determines the (types of) scales interacting with focus operators. The present study addresses this question by investigating the distributional behaviour of the additive scalar particle *even* relative to scales whose focus alternatives are ordered in terms of evaluative attitudes (positive, negative). Our hypothesis is that such evaluative attitudinal scales are at least partially functions of the lexical material in the sentential environment. This hypothesis is tested by determining correlations between sentence-level attitudes and lexically encoded attitudes in the relevant sentences. We use data from the Europarl corpus, a corpus of scripted and highly elaborated political speech, which is rich in argumentative discourse and thus lends itself to the study of attitudes in context. Our results show that there are in fact significant correlations between (manual) sentence-level evaluations and lexical evaluations (determined through machine learning) in the textual environment of the relevant operators. We conclude with an outlook on possible extensions of the method applied in the present study by identifying attitudinal patterns beyond the sentence, showing that positively and negatively connotated instances of *even* differ in terms of their argumentative function, with positive *even* often marking the climax and endpoint of an argument, while negative *even* often occurs in qualifying insertions like concessive parentheses. While we regard our results as valid, some refinements and extensions of the method are pointed out as necessary steps towards the establishment of an empirical sentence semantics, in the domain of scalar additive operators as well as more generally speaking.

1 Introduction

Scalar additive operators like Engl. *even*, *only*, Fr. *même*, *seulement*, Germ. *sogar*, *nur*, etc. are so called because they systematically interact with scales, i. e., sets of paradigmatic alternatives that are ordered in some way. In the simplest case, the focus values are numeric and there are entailment relations holding between the corresponding propositions, as in (1)

CC by 3.0

(here as well as in the following, the focus is enclosed by brackets with a subscript '$_F$', and focus alternatives are underlined, where present; *even* is rendered in italics).

(1) We are all familiar with examples of directives which remain in a vacuum for <u>five</u>, <u>ten</u> or *even* [fifteen]$_F$ years because they are not adopted by the Council. [Europarl][1]

Most cases are not as straightforward as (1), however, and the question of what orders scales interacting with focus operators has figured prominently in the theoretical literature on focus particles (cf. Jacobs 1983; Kay 1990; Löbner 1990; König 1991; Giannakidou 2007; Gast / van der Auwera 2011; among many others). While the 'classic' analyses (e. g. Karttunen / Karttunen 1977; Karttunen / Peters 1979; Rooth 1985) have regarded *even* and comparable operators as indicating that the focus represents a particularly unlikely option, in comparison to conceivable alternatives, many authors would agree today that scalar operators often do not interact with strictly semantic scales, and that the context has a considerable influence on their interpretation (cf. *inter alia* Fauconnier 1975; Anscombre / Ducrot 1983; Löbner 1990; see Gast / van der Auwera 2011 for a survey of relevant research, and Section 2.1 below for a brief summary).

If the interpretation of *even* and comparable operators is in fact context-dependent, the question arises what types of context features are responsible for determining the nature of the scales interacting with *even*. This question can only be answered by carrying out empirical studies of relevant operators in their textual environments, and this is precisely what the present study intends to do. Given the multi-dimensionality of the problem, and the more general difficulty of investigating matters of contextual embedding empirically, our study is restricted in some respects. We focus on the English additive scalar operator *even*, leaving scalar additive operators from other languages and restrictive operators like *only* for future research. Moreover, we restrict ourselves to one type of scale, i. e. evaluative attitudinal scales, where propositions are subjectively evaluated as (even) 'better' or (even) 'worse' from the speaker's point of view, in comparison to the focus alternatives under discussion.

The basic hypothesis underlying pragmatic analyses of *even* is that the relevant scales are determined by the context. The context comprises factors pertaining to one of (at least) two major sources of information (cf. Lewis 1972):

a. the situational (extra-linguistic) context, e. g. the coordinates of speech and shared (cultural, encyclopaedic) knowledge;

b. the linguistic context in the form of previous utterances made and linguistic material used (also often called 'co-text').

While we believe that both situational and linguistic context are relevant to the constitution of scales, we will focus on the linguistic context in the following, i. e., the co-text. More specifically, we will concentrate on the lexical material in the co-text. Our underlying hypothesis is that the lexical material in the co-text of an *even*-sentence has an influence on the type of scale that *even* interacts with. This hypothesis is operationalized as a hypothesis about correlations between lexically encoded connotations in the sentential environment of *even*, on the one hand, and proposition-level attitudes associated with *even*, on the other.

1 See Section 3.1 for some remarks on the Europarl corpus.

The study is structured as follows: In Section 2, we introduce the theoretical problem of determining scales interacting with *even*. The method of our study is described in Section 3. Section 4 presents the results with respect to correlations between lexical and sentence-level annotations. Section 5 contains an outlook on possible extensions of the method applied in the present study by considering the discourse environments of the sentences in question beyond the propositional level. Section 6 contains the conclusions.

2 The semantics of *even*: A brief overview

2.1 From semantic to pragmatic scales

Scales can be defined as sets of paradigmatic alternatives that are ordered in terms of some 'scaling dimension'. In 'traditional' analyses (e. g. Karttunen / Karttunen 1977; Karttunen / Peters 1979; Rooth 1985), *even* is analysed as implying or implicating that its focus is the most unlikely alternative from among the set of possibilities under discussion. This analysis explains why (2a) is natural, while (2b) is not. The latter sentence can only be interpreted by accommodating relevant background assumptions, e. g. that the speaker is not on good terms with his/her mother.

(2) a. *Even* [the Queen]$_F$ congratulated me on my birthday.
b. *#Even* [my mother]$_F$ congratulated me on my birthday.

The 'classic' unlikelihood analysis proposed by L. Karttunen and others has repeatedly been challenged, and it has been pointed out that scales interacting with the interpretation of *even*, while often in fact being scales of unlikelihood, are, to a certain extent at least, pragmatic, i. e., determined by the context (see for instance Jacobs 1983; Kay 1990; Löbner 1990; Rullmann 1997 for some discussion, and see Gast / van der Auwera 2010, 2011 for an overview with pertinent references). For reasons of space, we cannot reiterate the entire discussion of the semantic vs. pragmatic nature of *even*-scales here, and the reader is referred to the aforementioned literature for details. To illustrate the problem, we will just discuss one example showing that it is not always unlikelihood that orders scales interacting with *even*. Consider (3).

(3) [From a patent application for a '[f]lat element having a dark surface exhibiting a reduced solar absorption' (US Patent 7521118)]
It is more for aesthetic reasons that leather seats in automobiles are mainly coloured dark grey, indeed mostly *even* [black]$_F$.
Gast / van der Auwera (2011: 7)

Black is not a more unlikely colour for a car seat than dark grey. The reason that *even* is nevertheless felicitous in (3) is that what is at issue is not primarily the colour itself, but rather the consequences that the various colours (of car seats) have on the increase in temperature as a result of solar radiation. Black car seats absorb solar heat to a greater extent than grey ones do. In this particular case. The scale interacting with *even* can thus be represented as <maximal [black], sub-maximal [dark grey]> (absorption of solar radiation). The question arises how cases like (3) can be derived from a more general, context-sensitive analysis of *even*.

Adopting ideas developed by Carlson (1983), Roberts (1996, 2004) and Büring (2003), among others, Gast / van der Auwera (2011) have proposed an analysis according to which sets of alternatives associated with the focus of *even* are ordered by a dimension which is a function of the 'Quaestio' corresponding to a given utterance, i. e., the 'immediate question under discussion' (in terms of Roberts 1996, 2004; Büring 2003; the term 'Quaestio' has been adopted from Klein / von Stutterheim 1987). The Quaestio is the question to which a given sentence provides an answer. More often than not, the Quaestio of an utterance is implicit, and is assumed by the speaker to be accessible in the hearer's consciousness, or at least susceptible to accommodation.

In canonical cases like (2) above, the Quaestio corresponding to the utterance in question can be recovered directly from focus structure by replacing the focus with a *wh*-pronoun ('Who congratulated Fred?'). For other, more context-dependent, cases, Gast / van der Auwera (2011) have argued that the relation between an *even*-sentence and the corresponding Quaestio – and, hence, the relation between *even* and the scale that it interacts with – is more indirect. In cases such as (3), it is a contextual implication in the sense of Relevance Theory (Sperber / Wilson 1986) that provides an answer to the current Quaestio. This type of indirectness is illustrated in Figure 1 (a contextual implication is always determined relative to some assertion *A* and some Quaestio *Q*, and is therefore represented as I_C (*A*, *Q*)).

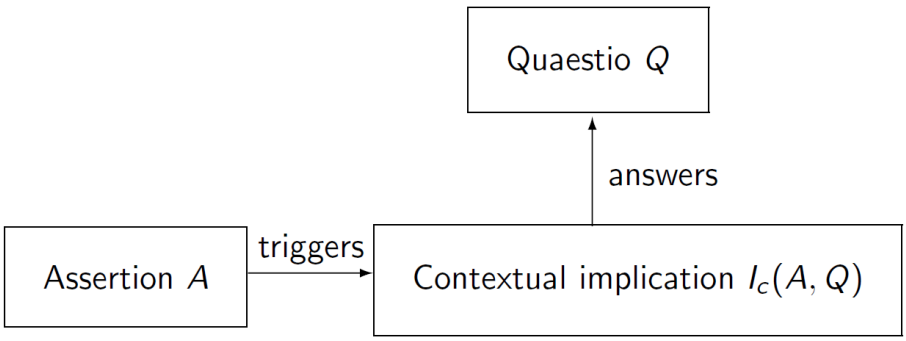

Figure 1: Indirect answers to a Quaestio.

In (3) above, there is a Quaestio relating to the degree of solar radiation that is absorbed, and the answer, while not providing direct information about this Quaestio, allows for an inference qualifying as a sufficient answer. The components of this particular instance of *even* can be summarized as shown in (4).

(4) a. Context
Automobile seats absorb solar heat.

b. Quaestio
How much heat do (specific types of) automobile seats absorb?

c. Assertion
Automobile seats are mostly dark grey or (even) black.

d. Contextual implications
Automobile seats absorb heat to a high though sub-maximal degree (dark grey), or even to a maximal degree (black).

The relationship between explicit focus alternatives, contextual implications and the scaling dimension ordering the sets of alternatives in (3) is shown in (5). Note that while there is no implicational relation holding between the properties 'dark grey' and 'black', the degrees of solar heat associated with each colour do form an implicational scale.

(5) Quaestio: To what extent is solar heat absorbed by car seats?

π_1	The seats are black	□	They absorb solar heat to *z* degree
π_2	The seats are grey	□	They absorb solar heat to *y* degree
…	…	□	…
π_n	The seats are white	□	They absorb solar heat to *a* degree

Ordering of focus values $z > y > \ldots > a$

Ordering of propositions $\pi_i \subset \pi_{i+1} \ldots \subset \pi_n$

Gast / van der Auwera (2011) have defined the notion of 'pragmatic strength' as a scaling dimension that orders sets of alternatives relative to some contextually given Quaestio (cf. also Anscombre / Ducrot's 1983 notion of 'argumentative strength'). Their definition is given in (6) (note that terms written in small caps are also explicitly defined in Gast / van der Auwera 2011).

(6) A proposition π is PRAGMATICALLY STRONGER (relative to a given Quaestio Q) than a proposition ϱ iff the RELEVANT CONTEXTUAL IMPLICATIONS of π (with respect to Q) entail the RELEVANT CONTEXTUAL IMPLICATIONS of ϱ (with respect to Q). (Gast / van der Auwera 2011: 9)

The notion of 'pragmatic strength', and the assumption that pragmatic strength orders alternatives contrasting with *even*, allows for a general analysis of the semantics of this particle, as it covers not only context-dependent examples like (3) above, but also more typical ones, e. g. cases of unlikelihood. While the occurrence of a more likely event does not imply the occurrence of a less likely event, the relevant probabilities – which can be retrieved via contextual implications – do stand in a relation of entailment to each other.

2.2 Attitude scales interacting with *even*

While the pragmatic analysis sketched in Section 2.1 can account for examples that are not easily explained in terms of a semantic analysis à la Karttunen/Karttunen/Peters, it does not say anything about attitudes held by the speaker towards a given state of affairs. Consider (7), dealing with the sad phenomenon of child abuse, which has been taken from the internet:

(7) The abuser, needless to mention, is not altogether a brute. He could be a relative, teacher, or *even*, as in most cases, a $[\text{parent}]_F$.[www][2]

The adverbial *as in most cases* shows that the scale is not one of unlikelihood, at least not of objective unlikelihood. It is of course conceivable to analyse this example in terms of perceived or subjective unlikelihood, or perhaps remarkability or "noteworthiness" (Herburger 2000). While the speaker uttering (7) is aware that parents are, sadly, not unlikely

2 http://libertarianeconomist.com/sexually-harassed, accessed on 10.07.2015.

abusers of their children, (s)he may still be aware that the reader does not have access to that information and therefore 'flag' the focus value 'parent' as particularly remarkable.

While the exact interpretation of notions such as 'unlikelihood' is certainly a non-trivial matter, and although unlikelihood can probably not be regarded as an objective phenomenon, it seems to be clear that in many cases, the use of *even* is at least partially motivated by the expression of an attitude, more specifically, an evaluative attitude. With respect to (7), for example, the sequence <relative, teacher, parent> suggests an ordering in terms of the degree of responsibility for, and authority over, children. Accordingly, there seems to be an element of evaluation involved, and the focus values could be regarded as being ordered in terms of 'despicability'. (8) (from the Europarl corpus) is another pertinent example.

(8) All these firms have gone ahead with job cuts and *even* [redundancies]$_F$. [Europarl]

In this case, we can probably recover a scale of likelihood, as redundancies are more unlikely than job cuts. At the same time, however, the expression of an attitude seems to be involved: In the context of parliamentary debate, redundancies are certainly evaluated even more negatively than job cuts.

Further examples of *even* that seem to carry attitudinal load are given in (9) and (10). In each case, the preceding context is provided to make the type of attitude expressed clearer.

(9) [I should add that today, at a time when Europe is revealing itself and is proving to be incapable of opposing the inclusion of Fascists in a government within the European Union, it is no longer acceptable to have an ultraliberalism which destroys public services foisted on us in the name of this same Europe.] I am a European and *even* [a federalist]$_F$. [Europarl]

(10) [This is not the way to fight extremism.] It may even be a way of [doing extremism a great deal of good]$_F$. [Europarl]

While the examples considered so far could probably also be analysed in terms of other scaling dimensions, there are instances of *even* that appear to be exclusively attitudinal, specifically when the focus itself is a basically evaluative expression, such as *worse* in (11).

(11) Today we have analysed the contract and found that the agreement with the EU not only creates the same sort of relationship as with NAFTA, but that, in parts, it is *even* [worse]$_F$. [Europarl]

Having pointed out the differences between pragmatic and semantic scales, and having identified evaluative attitudinal scales as a type of scale that is particularly relevant to the analysis of *even*, especially in the context of parliamentary debate, we will now turn to the question of how such attitudes can be investigated empirically.

3 Attitudes associated with *even*: An empirical study

Our study is based on the following general hypothesis:

(12) Evaluative attitudinal scales interacting with *even* are evoked by the context of the relevant sentence.

We take it that various context factors interact in the constitution of scales. In this study, we want to show that lexically encoded attitudes in the co-text (cf. Sect. 1) are one such factor. Accordingly, we can formulate the following, more specific, hypothesis:

(13) The presence of an evaluative attitudinal scale interacting with *even* correlates with attitudes conveyed by the lexical material in the sentential co-text.

The hypothesis in (13) is motivated by the assumption that pragmatic scales are a function of the Quaestio of an utterance (cf. Section 2.1). The Quaestio, in turn, can be expected to be reflected in the lexical material corresponding to the background (information-structurally speaking) of the relevant utterances.

In order to test the hypothesis in (13), we need a sample of attested examples which are annotated for evaluative attitudes at two levels, (i) at the propositional level, and (ii) at the lexical level. The processes of coding are described in Section 3.2 (for the proposition-level/manual annotations), and in Sections 3.3 and 3.4 (for the lexical/automatic annotations). Before proceeding to the process of data enrichment, we will make some remarks on the data itself in Section 3.1.

3.1 The corpus and the sample

The Europarl corpus contains the proceedings of the European Parliament (cf. Koehn 2005; Cartoni et al. 2013). It comprises approx. 60 million words per language (version 7, released in May 2012), and represents scripted political speech, a highly elaborated register in which lexical items relating to 'argumentation' (in the sense of Anscombre / Ducrot 1983) can be assumed to be used frequently and carefully. Moreover, the genre is characterized by a high degree of explicitness, which makes it easier for annotators to identify attitudes and sets of alternatives associated with a given focus than this would be the case in a corpus of, say, spontaneous face-to-face interaction.

For our study, the Europarl corpus has a second advantage: It is a translation corpus. We can thus use information from other languages to classify the instances of *even* in the English corpus part. Specifically, we have used the German corpus part as a 'filter': We extracted a sample of 200 randomly chosen examples of *even* which were rendered using the particle *sogar* into German. In this way we made sure that only uses of *even* with a comparable lexical semantics were included. As is well known, *even* is highly polysemous and corresponds to a large number of translation equivalents in languages that make more lexical differentiations in the domain of focus particles, like, for instance, German (cf. König 1982). The sample was of course also manually inspected. It contains both original language and translations, which is irrelevant, as far as we can tell, as it seems to us that lexical connotations are not subject to any significant translation effects between German and English (we are not aware of any empirical studies of that matter, however). Our sample (with the annotations) is publicly available on the following URL: http://www.uni-jena.de/~mu65qev/data (accessed on 14.07.2015).

3.2 Manual annotation: Attitudes in the scope of *even*

In a first step, the data was coded manually for evaluative attitudes conveyed in the scope of *even*. As an operational test, we used the insertion of adverbials expressing either a positive or a negative attitude (see for instance Bonami / Godard 2008 on evaluative adverbs). If such adverbials could be inserted without noticeably changing the meaning of the proposition, they were regarded as merely 'echoing' an attitude which was there independently.

For veridical propositions, i. e. propositions occurring in a context in which they are implied to be true (cf. Zwarts 1995), the adverbs *unfortunately* and *sadly* were used as diagnostics for a negative attitude, and *fortunately* for positive attitudes. Sometimes the wider context had to be inspected in order to identify a speaker's attitude. The negative adverbs *unfortunately* and *sadly* were taken to differ in 'strength' and the degree of personal involvement.

Let us consider some examples. (14) was classified as 'Negative', as the insertion of *sadly* is fully compatible with the content of the sentence (diagnostic adverbials are rendered in bold face in the following).

(14) There is no way out for victims and the very nature of the environment in which they work often reduces them to <u>despair and drug addiction</u>. [**Sadly**], [s]ome *even* go as far as [committing suicide]$_F$. [Europarl]

In order to distinguish (manual) scope-level annotations – which we associate with *even* – from (automatically determined) lexical annotations, we will use the subscripts '*even*' and '*lex*' in the following. Examples like (14) will thus be classified as 'N*even*'.

Some sentences with *even* contained a diagnostic adverb already and were thus also classified as 'N*even*':

(15) **Unfortunately**, and as everyone recognised during the military action in Kosovo, there were civilian sectors, and in particular infrastructural <u>sectors</u> directly affecting <u>the general public</u>, and *even* [inhabited areas]$_F$, which were directly affected by NATO air attacks. [Europarl]

Two examples of positively evaluated propositions ('P*even*') are given in (16) and (17).

(16) Turning to the report's proposals to the House, I am very pleased to note that all of the budgetary objectives established in previous programmes have [**fortunately**] <u>been met,</u> and *even* [exceeded]$_F$. [Europarl]

(17) All these instruments are currently being examined by the Council. [**Fortunately**,] [e]ven [the United Kingdom and Ireland]$_F$ have decided to join the Member States on this and Denmark, which does not have the same capacity to participate in work on civil legal cooperation, is also seeking solutions which are currently being examined. [Europarl]

For non-veridical propositions (propositions that are not implied to be true), the adverbials *in the worst case* and *ideally* were used as diagnostics. Two N*even*-examples are given in (18) and (19), two P*even*-examples in (20) and (21).

(18) Delay in justice can mean personal difficulties and [**in the worst case**] *even* [tragedy]$_F$. [Europarl]

(19) This is not the way to fight extremism. It may [**in the worst case**] *even* be a way of [doing extremism a great deal of good]$_F$. [Europarl]

(20) Mr President, the food problem can, in principle, be solved, [**ideally**] *even* in [Africa]$_F$, provided we have a cohesive development policy intended to reduce poverty there instead of increasing wealth over here. [Europarl]

(21) I would remind you of the commitment given by Member States in relation to the world's highly indebted poor countries to reduce and [**ideally**] *even* [cancel their debt]$_F$ [...] [Europarl]

If the diagnostic adverbials listed above could not be inserted without introducing an attitude which was not recoverable from the 'bare' sentence, or if the relevant sentences were just infelicitous with any such adverbial, they were classified as 'O(bjective)'. Two pertinent examples are given in (22) (veridical) and (23) (non-veridical).

(22) We do drink port wine, and we do like German beer, and [#**fortunately**/#**unfortunately**] we *even* use Finnish saunas, but it is not from Europe that culture is under threat. [Europarl]

(23) The noise impact [#**fortunately**/#**unfortunately**] differs even within the same airport according to the runway used. For example, runway 25 of Fiumicino airport is next to the sea and hardly disturbs anybody at all, whereas if an aircraft were to take off from the right-hand section of runway 16 it would disturb half a million people or so from the Ostia and Fiumicino communities. [Europarl]

Before considering correlations between proposition-level evaluations and lexical connotations let us have a look at the frequencies of the three types of attitudinal values in our sample. Our sample contains 92 observations of examples classified as N$_{even}$ (46%), 36 observations of P$_{even}$-examples (18%) and 72 of category O$_{even}$ (36%). It is of course likely that the distribution of positive and negative attitudes in the sample is at least partially specific to the register of political speech. Future studies will have to show to what extent the bias towards negativity observed in our data is found in other corpora as well.

3.3 Lexical connotations

In order to automatically identify attitudes associated with the lexical material in *even*-sentences, we used the SentiWordNet database (cf. Esuli / Sebastiani 2006; Baccianella et al. 2010), which assigns 'sentiment scores' to WordNet synsets. There are three sentiment scores for each synset, 'N(egative) ', 'P(ositive) ' and 'O(bjective) '. N, P and O add up to 1 (N + P +O =1); N + P can also be regarded as an indicator of 'subjectivity'.

The sentiment scores of the SentiWordNet database were determined through semi-supervised machine learning. For reasons of space, we cannot discuss the details of the method here, and the reader is referred to Esuli / Sebastiani (2006) and Baccianella et al. (2010), as well as references cited there. The method "relies on training, in a semi-supervised way, a binary

classifier that labels terms as either Positive or Negative. A semi-supervised method is a learning process whereby only a small subset $L \subset Tr$ of the training data Tr are human-labelled. In origin the training data in $U = Tr - L$ are instead unlabelled; it is the process itself that labels them, automatically, by using L (with the possible addition of other publicly available resources) as input." (Esuli / Sebastiani 2006: 196)

The human contribution to this process is rather minimal and hardly goes beyond the decision to use the adjectives *good* and *bad*, as well as synonyms and indirect antonyms, as the basis for the machine learning algorithm. The glosses of WordNet served as a textual basis for the process.

It should be noted that most 'subjective' words have both N- and P-scores. This is reasonable, considering that basically positive words can occur in negative contexts, and vice versa, e. g. under negation and in other non-veridical contexts. A screenshot of a basically positively connotated synset – 'honourable' – is shown on top in Figure 2, a negatively connotated example ('criminal') underneath.

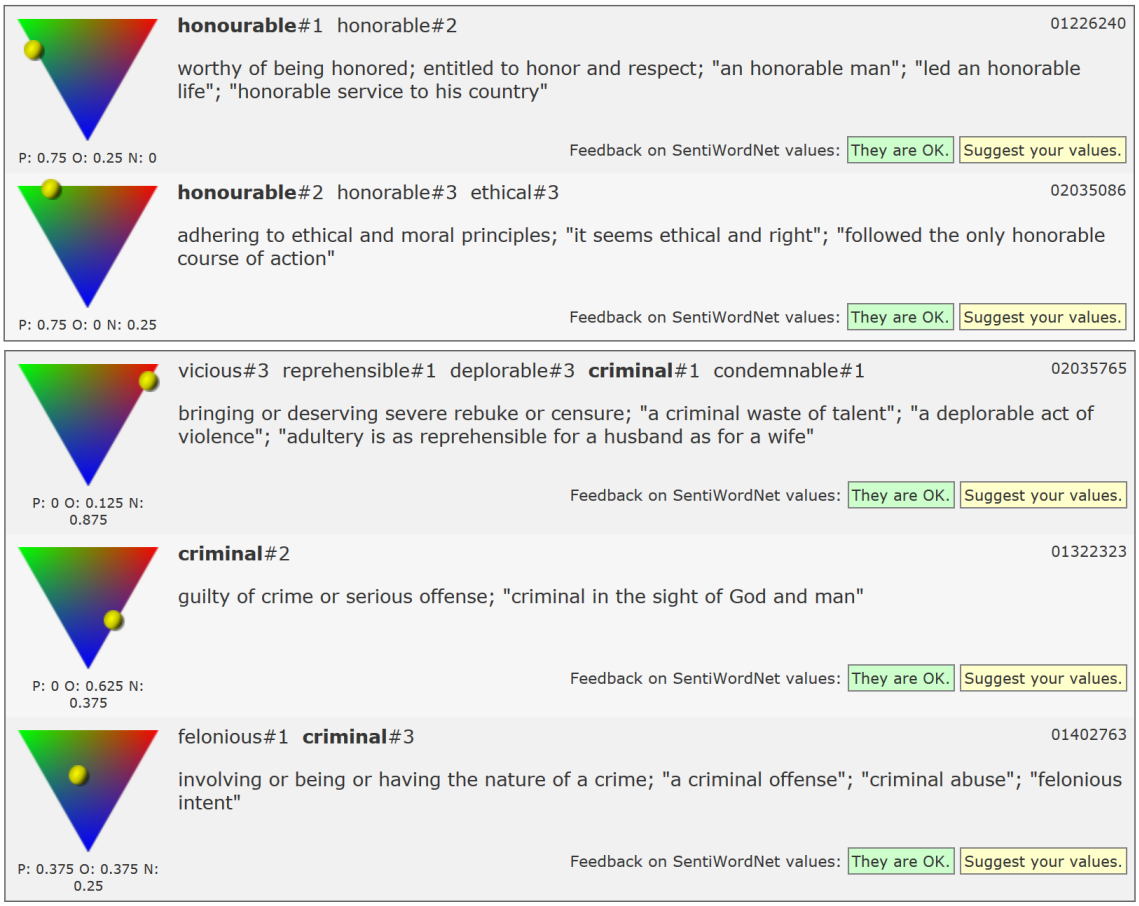

Figure 2: The sentiment scores of *honourable* and *criminal*.

The triangular symbolic representations in Figure 2 distinguish between the dimension 'subjective vs. objective' on the vertical axis, and between 'positive' and 'negative' attitudes on the horizontal axis. The more objective a term is, the less the distinction between positive and negative evaluations plays a role.

As far as the linguistic interpretation of the sentiment scores is concerned, Esuli / Sebastiani (2006: 418) (Note 1) point out that "[...] associating a graded score to a synset for a certain

property (e. g. Positive) may have (at least) three different interpretations: (i) the terms in the synset are Positive only to a certain degree; (ii) the terms in the synset are sometimes used in a Positive sense and sometimes not, e. g. depending on the context of use; (iii) a combination of (i) and (ii) is the case. Interpretation (i) has a *fuzzy* character, implying that each instance of these terms, in each context of use, have the property to a certain degree, while interpretation (ii) has a *probabilistic* interpretation (of a frequentistic type), implying that membership of a synset in the set denoted by the property must be computed by counting the number of contexts of use in which the terms have the property. We do not attempt to take a stand on this distinction, which (to our knowledge) had never been raised in sentiment analysis and that requires an in-depth linguistic study, although we tend to believe that (iii) is the case."

The question arises how sentiment scores are interpreted and what type of variable they represent in an empirical study. The values are minimally ordered, i. e. they can be regarded as an ordinal variable. It seems to be clear that a word with a P-score of 0.5 is evaluated more positively than a word with a P-score of 0.25. But is the 0.5-word 'twice' as positive as the 0.25-word? This would hold if sentiment scores were ratio-scaled, or probabilistic, in terms of the above quotation. Given the obvious limitations of the method, we will take a conservative stance and regard sentiment scores as an ordinal variable.

3.4 Quantifying lexical connotations

In order to determine and quantify the lexical connotations associated with a given co-text on the basis of the SentiWordNet database, we had to prepare the textual material in certain ways. After a clean-up with regular expressions and another round of manual inspection of the 200 examples, the corpus was first POS-tagged using an R-implementation of the Apache openNLP-tagger.[3]

In a second step, each content word was assigned to a specific WordNet synset, i. e., the words were disambiguated. This was done using a slightly customized version of the Perl module SenseRelate.[4] SenseRelate operates on the basis of word contexts or word windows to associate WordNet senses with words in particular contexts. The authors of the module describe the algorithm behind SenseRelate's word sense disambiguation in detail in Pedersen et al. (2005).

Consider (24) for illustration:

(24) Mr President, first of all, we want to denounce the fathomless hypocrisy of those who claim to be concerned about the environment and water, but whose criminal activities such as the attacks against Yugoslavia, aside from leaving thousands dead and wounded, have also brought about huge ecological disasters for water resources too, making them not only unusable but *even* [extremely hazardous]$_F$. [Europarl]

The result of the automatic annotation process as described above is shown in (25). The tokens are followed by a POS-tag, separated by #, and followed by the number of the WordNet synset corresponding to the relevant meaning.

3 https://opennlp.apache.org/, http://cran.r-project.org/web/packages/openNLP/index.html, both accessed on 10.07.2015.

4 http://www.d.umn.edu/~tpederse/senserelate.html, accessed on 10.07.2015.

(25) Mr#n#1 President#n#2 first#r#2 of#r#ND all#CL we#CL want#v#1 to#CL denounce#v#2 the#CL fathomless#a#ND hypocrisy#n#2 of#r#ND those#CL who#CL claim#v#1 to#CL be#v#1 concern#v#1 about#r#3 the#CL environment#n#1 and#CL water#n#1 but#CL whose#CL criminal#a#3 activity#n#5 such#a#1 as#r#1 the#CL attack#n#1 against#r#ND Yugoslavia#n#1 aside#r#2 from#r#ND leave#v#5 thousand#n#1 dead#a#1 and#CL wound#v#1 have#v#2 also#r#1 bring#v#1 about#r#4 huge#a#1 ecological#a#1 disaster#n#2 for#r#ND water#n#2 resource#n#2 too#r#1 make#v#5 them#CL not#r#1 only#r#2 unusable#a#1 but#CL even#r#3 extremely#r#1 hazardous#a#1.

Note that the words are lemmatized, as can be seen, for instance, from 'concern#v#1', corresponding to the inflected form *concerned* in the original. For our purposes, SenseRelate was modified to immediately include lemma information together with the word sense disambiguation.

The format in (25) could be used to determine sentiment scores for each word in the corpus. (26) shows the 'N'-values for all words in (24) that are categorized in the SentiWordNet database.

(26) hypocrisy#n#2#0.375 criminal#a#3#0.25 activity#n#5#0.125 such#a#1#0.125 dead#a#1#0.75 wound#v#1#0.5 huge#a#1#0.125 ecological#a#1#0.375 disaster#n#2#0.5 unusable#a#1#0.625 hazardous#a#1#0.125.

In this way, we determined sentiment scores not just for the *even*-sentences themselves, but also for the surrounding context, i. e., the preceding and following sentence. In the following, we will refer to the surrounding sentences as the 'pre-text' and the 'post-text' of an *even*-sentence.

Let us consider the general distribution of sentiment scores, focusing on the *even*-sentences themselves. Our sample contained a total of 6605 words (in the *even*-sentences, i. e., without pre-text and post-text). Of these, 495 carry a positive sentiment score, 405 a negative one. The large majority of words do not have a sentiment score. Figure 3 shows the frequencies of P*lex*- and N*lex*-scores in our sample.

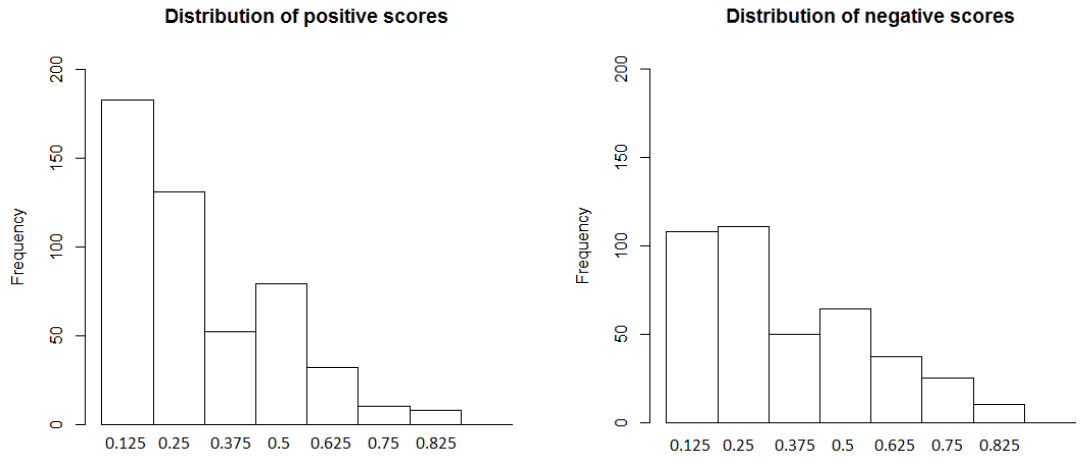

Figure 3: Distribution of sentiment scores, positive (left) and negative (right).

As Figure 3 shows, positive and negative sentiment scores are distributed differently. For the negative scores (on the right), the most frequent category is that of 0.25 – though it is only slightly more frequent than the category of 0.125 – while for the positive scores, 0.125 is the most frequent category. The different overall distributions of P*lex*- and N*lex*-scores in general might at least partially be responsible for the differences between P*even*- and N*even*-sentences to be discussed in Section 4.

4 Testing proposition-level and lexical annotations for correlations

The SentiWordNet-scores can obviously only provide rough approximations to attitudes expressed in natural language, for two reasons. First, the method itself has limitations, relying as it does on probabilistic patterns of cooccurrence, which can approximate, but never categorically identify, attitudes. Second, the degree of attitudinal information conveyed by lexical items is limited, as lexical meanings, in general, interact closely with the context. It is for this reason that many words have both positive and negative sentiment scores (cf. Section 3.3).

Remember that the two types of annotations used for the present study – manual and automatic ones – are located at different levels of analysis. Manual annotations are properties of sentences or propositions (as the semantic correlates of sentences), while automatic annotations are properties of words. What we can correlate, thus, is the lexical sentiment scores P*lex* and N*lex* occurring in a given sentence, and the classifications of the relevant instances of *even* as P*even*, O*even* or N*even*.

Let us start by considering absolute frequencies. Table 1 shows the frequencies of P*lex*- and N*lex*-scores in sentences of categories P*even* and N*even*.

	prop-level	0.125	0.25	0.375	0.5	0.675	0.75	0.875
P_{lex}	P_{even}	33	30	10	22	6	3	2
	O_{even}	67	46	15	24	16	3	3
	N_{even}	83	55	27	33	10	4	3
N_{lex}	P_{even}	15	28	10	10	5	3	0
	O_{even}	36	37	15	19	8	5	1
	N_{even}	57	46	25	35	24	17	9

Table 1: Frequencies of P_{lex}- and N_{lex}-scores in examples of category P_{even}, O_{even} and N_{even}.

The upper half of Table 1 is visualized in Figure 4 in the form of a Cohen-Friendly association plot (cf. Cohen 1980, Friendly 1992). Note that no significance levels are represented here. We will return to the statistics below. What matters for the time being is the fact that the boxes on the top line, corresponding to sentences manually classified as positive (P*even*), are predominantly overrepresented in the right half – as is indicated by their position above the baseline – while they are underrepresented/underneath the baseline on the left (note that the size of a box is proportional to the deviation from statistical independence). A different pattern can be observed for the boxes on the bottom line, corresponding to the positive scores in sentences manually classified as negative (N*even*), where only the lower scores (on the left) show (some) overrepresentation. The scores for the O*even*-category are rather heterogeneous.

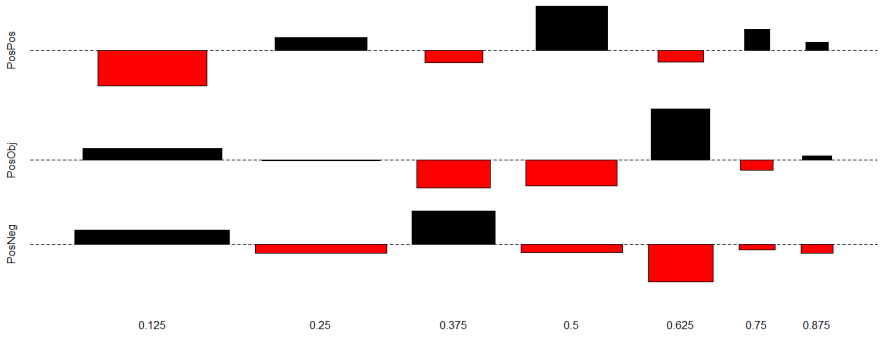

Figure 4: The distribution of P_{lex}-scores relative to proposition-level annotations.

The distribution of negative scores relative to the three types of sentences is shown in Figure 5. The picture is clearer here than in Figure 4. The top row – corresponding to P*lex*-scores in N*even*-sentences – shows a steady decrease from 0.25 to 0.875 (though there is a sharp increase from 0.125 to 0.5), while the bottom row – showing the distribution of N*lex*-scores in N*even*-sentences –monotonously increases from 0.25 to 8.75. O*even*-sentences seem to pattern more with those of category P*even*.

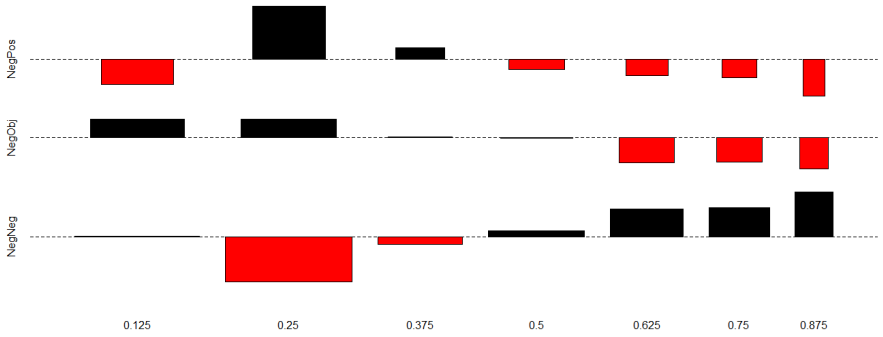

Figure 5: The distribution of N_{lex}-scores relative to proposition-level annotations.

In order to estimate degrees of statistical significance, we fitted a logistic regression model, treating the sentiment scores as an ordinal variable (cf. Section 3.3). Again, the results are strikingly different for positive and negative scores. Let us start with the positive scores. The models were fitted using the *lrm* ()-function of the *rms*-package for R (R Core Team 2013). The models are intended to predict for any given word with a specific (positive or negative) sentiment score whether or not it will occur in a sentence manually classified as 'P*even*'. The response variables have thus been binarized to P_{even} vs. {$N_{even,}$ O_{even}}, or N_{even} vs. {P_{even}, O_{even}}. The predictor has seven ordered levels, <1.25, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875>.

The model statistics are very poor (LR X^2 =4.04, df=6, p=0.67). The model is represented graphically in Figure 6. The little triangles represent the odds ratios on a logarithmic scale, and the lines show confidence intervals. The second thickest line corresponds to the %95-level. The further left a triangle and the surrounding confidence interval is located on the

diagram, the more closely it is associated with the category P*even*. 0.125 is the reference level.

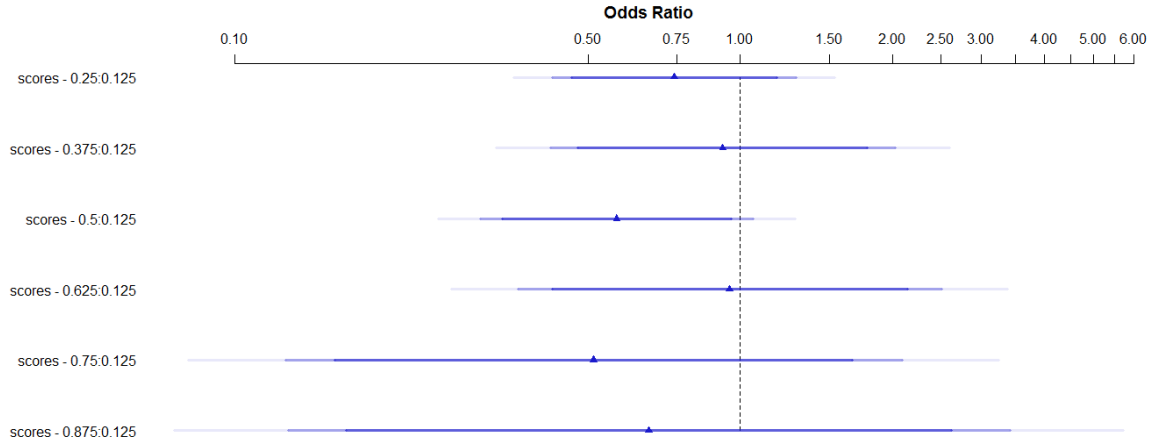

Figure 6: Logistic regression model for P_{lex}-annotations in P_{even}-sentences.

The model for the negative sentiment scores looks entirely different. The likelihood ratio test statistics are much better than for the positive scores (LR X^2 =17.07, df=6, p=0.009). Figure 7 provides a graphical representation of the model. It clearly suggests a correlation between the lexical and the proposition-level annotations.

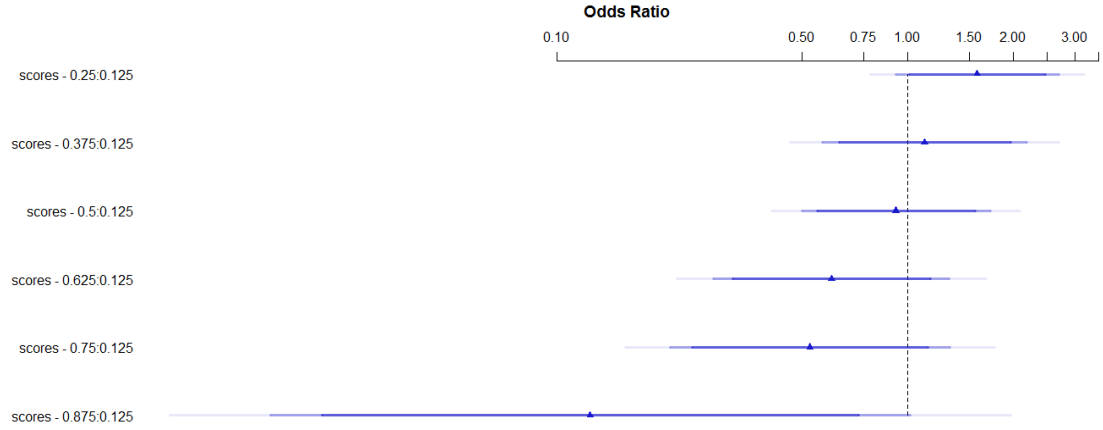

Figure 7: Logistic regression model for N_{lex}-annotations in N_{even}-sentences.

The high degree of uncertainty, reflected in large confidence intervals, of both models is obviously, at least partly, due to scarcity of data. Clearly, more comprehensive investigations and larger datasets are needed for more reliable predictions. For the time being, it seems reasonable to lump some of the data together. This makes sense from a theoretical point of view anyway. Remember that positive and negative sentiment scores are not mutually exclusive, i. e., many words exhibit non-zero values for both categories. We can thus distinguish between those scores that indicate a 'prevailingly positive' evaluation and those with a 'prevailingly negative' orientation. A word can be assumed to be prevailingly positive or negative when its sentiment score is ≥ 0.5. Words of this class will be said to have a 'high' P- or N-score, while the others will accordingly be called 'low'.

The resulting models are more accurate than the ones based on eight levels. High P_{lex} or N_{lex}-scores are significant predictors for the occurrence in sentences of category P_{even} or N_{even}. The model statistics are shown in Table 2. For both positive and negative scores, the category 'high' exhibits a p-value < 0.05. Unsurprisingly, the value for negative scores (p=0.0002) is

much lower than the one for positive scores (p=0.0226). The models are represented in the same graphical format as used above in Figure 8.

	Coef	S.E.	Wald Z	Pr(> /Z /)
P_{lex} in P_{even}				
intercept	1.7570	0.3063	5.74	<0.0001
low	0.0471	0.1245	0.38	0.7054
high	-0.4355	0.1909	-2.28	0.0226
N_{lex} in N_{even}				
intercept	0.4954	0.2135	2.32	0.0203
low	0.0908	0.1074	0.85	0.3977
high	-0.6865	0.1868	-3.67	0.0002

Table 2: Binary logistic regression models with two predictor variables.

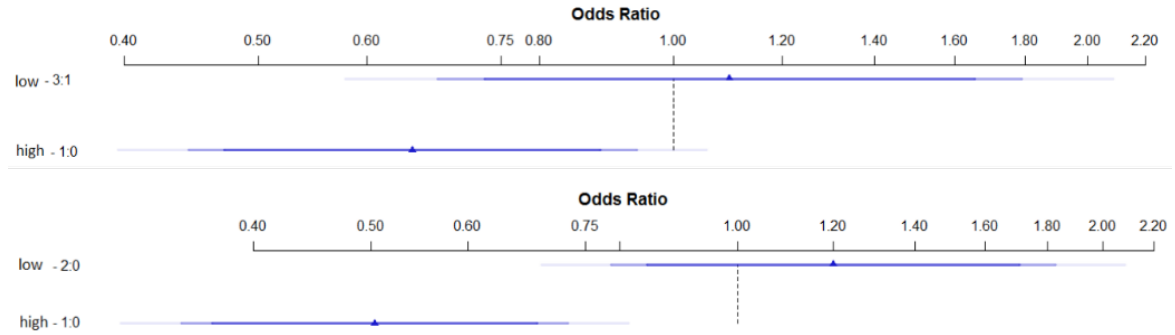

Figure 8: Binary logistic regression models distinguishing only two categories of scores (top: P_{lex} in P_{even}; bottom: N_{lex} in N_{even}).

We can conclude that both P*lex*- and N*lex*-scores correlate significantly with sentence-level annotations if they are binned into the two classes 'low' ($0 \leq x_{low} < 0.5$) and 'high' ($x_{high} \geq 0.5$). Future, more comprehensive studies will have to show if sentiment scores from the SentiWordNet-database can also be used as predictors at a higher level of granularity.

5 An outlook: *Even*-sentences in their textual environments

Our primary goal has been to test whether proposition-level attitudes correlate with lexical attitudes in *even*-sentences. In this section, we would like to show that the method that we have applied opens up new avenues for the study of focus operators in their textual environments. We can investigate, for instance, variation in attitude scores, i. e., differences between the scores of an *even*-sentence and the preceding as well as following sentences. As we will show in this section, this approach can bring to light interesting differences between positively and negatively evaluated instances of *even*, as far as their argumentative function is concerned.

Let us first consider variation in the occurrence of words with high P*lex*-scores, in the environment of P*even*-sentences. Table 3 shows the mean values of the number of words with a high P*lex*-score in the environment of P*even*- sentences, as opposed to those of category N*even* and O*even*. These scores are represented diagrammatically in Figure 9.

	pre-text	*even*-sentence	post-text
P_{even}	0.92	0.97	0.5
N_{even}, O_{even}	0.60	0.58	0.6

Table 3: High-score P*lex*-words with pre- and post-text (mean values).

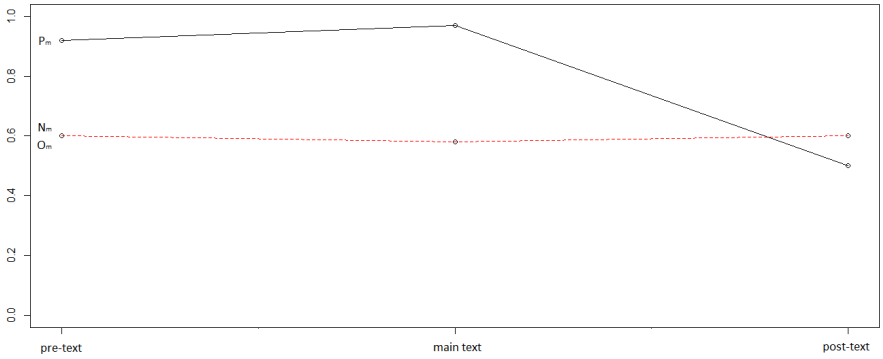

Figure 9: High-score P*lex*-words.

The values for non-P*even*-sentences are rather stable from the pre-text to the post-text, showing almost no variation. The P*even*-sentences show no significant change between the pre-text and the *even*-sentence (p=0.77, according to a t-test), but there is a significant drop from the main text to the post-text (p=0.031).

The situation for the N*lex*-scores is quite different (cf. Table 4 and Figure 10). Their distribution relative to non-N*even*-sentences is stable, more or less as in the case of P*lex*-scores in non-P*even*-sentences. The slight increase from the pre-text to the main text (0.39 □ 0.45) is not significant (p=0.5). However, the changes from the pre-text to the main text, and from the main text to the post-text, in N*even*-examples, are significant (p=0.05 for pre-to-main, and p=0.006 for main-to-post).

These results show that positive sentiment scores in the context of examples of type P*even*, and negative scores around examples of type N*even* , are distributed quite differently. N*even*-examples are typically N*lex*-'peaks', in the sense that their N*lex*-scores differ significantly from the scores of the preceding and following sentences. The P*lex*-scores in P*even*-examples, by contrast, suggest that positive attitudes are conveyed even before the *even*-sentence.

	pre-text	*even*-sentence	post-text
N_{even}	0.64	0.92	0.55
P_{even}, O_{even}	0.39	0.45	0.45

Table 4: High-score N$_{lex}$-words with pre- and post-text (mean values).

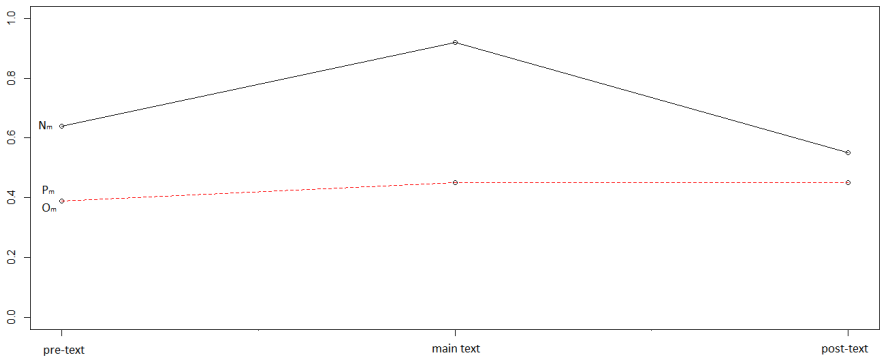

Figure 10: High-score N*lex*-words.

While the results are statistically robust, we do not know at present what exactly this finding entails. What it suggests is the following: While there seems to be a tendency for positively connotated instances of *even* to mark the climax of an argument, negative instances of *even* are seemingly often insertions, in terms of argumentation, for instance, concessive or qualifying parentheses, or remarks referring to a possible counterargument. Consider the discourse sequence in (27) (the words with a sentiment score categorized as 'high' are printed in bold face).

(27) a. However, this current **satisfactory** situation must not mask a series of shortcomings and must lead to the implementation of the series of recommendations which the rapporteur, Mr Pomés Ruiz, correctly makes in his report.

b. He mentioned **one** of these recommendations in his speech: the need to take **advantage** of the current favourable situation, the **healthy** state of public finances – which on the other hand are disciplined by the **stability** and convergence programmes – in order to reduce the deficit, and possibly remove it completely, as well as to reduce debt levels, *even* to [below the limits set in the convergence programmes]$_F$. [Europarl]

The *even*-sentence in (27b) continues the argument made in (27a), which reflects the speaker's position. The main point made here is that the "current satisfactory situation" should be consolidated through the "implementation of the series of recommendations" made by Mr. Pomés Ruiz. This claim is supported in the next sentence, where it is pointed out that the deficit as well as the debt levels can be reduced, "possibly [...] even below the limits set in the convergence programmes". (27b) provides an argument supporting the claim made in (27a).

Let us now consider a typical example of negative/parenthetical *even*. There is no word with a high negative sentiment score in (28a), which explicitly reflects the speaker's position. (28b) contains two words with a high negative sentiment score (and the rest of the sentence also carries prevailingly negative connotations). It is a concessive insertion, qualifying the previous statement somewhat. In (28c), the speaker returns to the positive main thread again, and the sentence does not contain a single word with a negative sentiment score.

(28) a. We are all aware of the enormous expectations of our people with regard to freedom, security and justice, particularly social justice.

b. Yet their lack of interest and involvement and sometimes even their **distaste** for all things political requires us to take specific action to tackle their **problems**.

c. This is the sine qua non condition for reconciling the popular and political spheres. [Europarl]

The examples in (27) and (28) have obviously been selected as typical representatives illustrating the tendency for P*even*-sentences to mark the climax of a positive argument, whereas N*even*-sentences seemingly often tend to occur in insertions, e.g. in qualifications or concessions. More data, and other types of annotations are obviously needed to test whether this tendency applies at a more global level. It is very likely, of course, that the tendency

observed in this section is at least partially register-specific, an assumption which can only be tested by considering data from a broader range of corpora.

6 Conclusions

The starting point of our investigation was the hypothesis that scales interacting with *even* (and, most likely, other types of operators) are pragmatic and thus (at least partially) determined by the context. We have focused on one type of scale and context feature, i. e., evaluative attitudes. The specific hypothesis that we have tested says that the evaluation of an *even*-sentence at the propositional level should be reflected in the lexical connotations conveyed in the sentential environment. In other words, we have hypothesized that lexically encoded connotations (as recorded in the SentiWordNet database) are significant predictors for the proposition-level classification of an instance of *even* as 'positive' or 'negative'. This hypothesis has been confirmed for a binary binning of the eight factors provided by the SentiWordNet database. The presence (and number) of words with a high P*lex*- or N*lex*-score turned out to be a significant predictor of the evaluative orientation of the sentence in question.

We have also aimed to show that the use of sentiment scores opens up new possibilities for the study of attitudes in discourse context. By determining scores in the discourse environment of *even*-sentences we were able to identify differences in the textual distribution of positively and negatively oriented instances of *even*. P*even*-sentences seem to mostly continue an argument and mark its endpoint and climax, wheras N*lex*-sentences often 'stick out' from the context and mark individual peaks. They typically seem to occur in insertions, e. g. in concessive qualifications.

While we believe that our results are valid, it has become clear that the study of sentential operators in context actually requires a broader range of methodological means than we have used for the present study. First, it seems to us that better results could be obtained by distinguishing between sentiment scores in the scope of a given example, as opposed to those outside its scope. In order to take this factor into account, we would need to annotate our data for scope relations. Second, the presence and nature of focus alternatives in the immediate discourse environment has not been taken into account in our study. They can obviously be expected to be good predictors of the evaluative orientation of an *even*-sentence. Then again, we would first have to annotate our sample for this feature. Finally, the SentiWordNet database, having proven a very useful tool for our study, has obvious limitations, too, and other methods or resources might give us an even clearer picture of the textual distribution of attitudes relative to scalar operators. We intend to tackle these and related challenges in the near future, in this way hoping to make a contribution to the establishment of an empirical – corpus-based and quantitative – study of sentence semantics.

References

Anscombre, Jean-Claude/Ducrot, Oswald (1983): *L'argumentation dans la langue*. Brussels: Pierre Mardaga.

Baccianella, Stefano et al. (2010): "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In: Calzolari, Nicoletta et al. (eds.): *Proceedings*

of the 7th International Conference on Language Resources and Evaluation. Valletta, ELRA: 2200–2204.

Bonami, Olivier/Godard, Danièle (2008): "Lexical semantics and pragmatics of evaluative adverbs". In: McNally, Louise/Kennedy, Christopher (eds.): *Adverbs and Adjectives: Syntax, Semantics, and Discourse*. Oxford, Oxford University Press: 274–304.

Büring, Daniel (2003): "On D-Trees, Beans, and B-Accents". *Linguistics and Philosophy* 26: 511–545.

Carlson, Laurie M. (1983): *Dialogue games: An approach to discourse analysis*. Dordrecht: Reidel.

Cartoni, Bruno et al. (2013): "Using the Europarl corpus for cross-linguistic research". *Belgian Journal of Linguistics* 27: 23–42.

Cohen, A. (1980): "On the graphical display of the significant components in a two-way contingency table". *Communications in Statistics – Theory and Methods* A9: 1025–1041.

Esuli, Andrea/Sebastiani, Fabrizio (2006): "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining". In: *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, Genova, IT: 417–422.

Fauconnier, Gilles (1975): "Pragmatic scales and logical structures". *Linguisic Inquiry* 6: 353–375.

Friendly, M. (1992): "Graphical methods for categorical data". In: *Proceedings of the SAS User Group International Conference* 17: 1367–1373.

Gast, Volker/van der Auwera, Johan (2010): "Vers une typologie des opérateurs additifs scalaires". In: Hadermann, Pascale/Inkova, Olga (eds.): *Approches de la scalarité*. Genève, Droz: 285–311.

Gast, Volker/van der Auwera, Johan (2011): "Scalar additive operators in the languages of Europe". *Language* 87: 2–54.

Giannakidou, Anastasia (2007): "The lanscape of *even*". *Natural Language and Linguistic Theory* 25: 39–81.

Herburger, Elena (2000): *What counts: Focus and quantification*. Cambridge, MA: MIT Press.

Jacobs, Joachim (1983): *Fokus und Skalen: Zur Syntax und Semantik der Gradpartikeln im Deutschen*. Tübingen: Niemeyer.

Karttunen, Fances/Karttunen, Lauri (1977): "Even questions". In: Kegl, Judy A. et al. (eds.): *Proceedings of the 7th meeting of the North Eastern Linguistic Society*. Cambridge, MA, MIT Press: 115–34.

Karttunen, Lauri/Peters, Stanley (1979): "Conventional implicature in Montague Grammar". In: Oh, Choon-Kyu/Dinneen, David A. (eds.): *Syntax and semantics, Vol. 11: Presuppositions*. New York, Academic Press: 1–56.

Kay, Paul (1990): "Even". *Linguistics and Philosophy* 13: 59–111.

Klein, Wolfgang/Stutterheim, Christiane von (1987): "Quaestio und referenzielle Bewegung in Erzählungen". *Linguistische Berichte* 108: 163–183.

Koehn, Philipp (2005): "Europarl: A parallel corpus for statistical machine translation". *MT Summit* 5: 79–86.

König, Ekkehard (1982): "Scalar particles in German and their English equivalents". In Lohnes, Walter F. W./Hopkins, Edwin A. (eds.): *The Contrastive Grammar of English and German*. Ann Arbor, Karoma Publishers: 76–101.

König, Ekkehard (1991): *The Meaning of Focus Particles*. London: Routledge.

Lewis, David (1972): "General semantics". In: Davidson, Donald/Harman, Gilbert (eds.): *Semantics for Natural Language*. Dordrecht, Reidel: 169–218.

Löbner, Sebastian (1990): *Wahr neben Falsch: Duale Operatoren als die Quantoren natürlicher Sprache*. Tübingen: Niemeyer.

Pedersen, Ted et al. (2005): "Maximizing Semantic Relatedness to Perform Word Sense Disambiguation". *Technical report, University of Minnesota Supercomputing Institute*. Research Report UMSI 2005/25.

R Core Team (2012): *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Wien: R Foundation for Statistical Computing. http://www.R-project.org, accessed 09.07.2015.

Roberts, Craige (1996): "Information structure in discourse: Towards an integrated formal theory of pragmatics". In: *Working Papers in Linguistics–Ohio State University Department of Linguistics*: 91–136.

Roberts, Craige (2004): "Context in dynamic interpretation". In: Horn, Laurence/Ward, Gregory (eds.): *The Handbook of Pragmatics*. London, Blackwell: 197–220.

Rooth, Mats (1985): *Association with Focus*. Doctoral Dissertation, University of Massachusetts.

Rullmann, Hotze (1997): "Even, polarity, and scope". In: Wiebe, Grace et al. (eds.): *Papers in Experimental and Theoretical Linguistics*. Department of Linguistics, University of Alberta: 40–64.

Sperber, Dan/Wilson, Deirdre (1986): *Relevance: Communication and Cognition*. Oxford: Blackwell.

Zwarts, Frans (1995): "Nonveridical contexts". *Linguistic Analysis* 25: 286–312.