Language and Cognition **12** (2020), 343–359. doi:10.1017/langcog.2019.50

Ranking Dutch intensifiers: a usage-based approach

MICHAEL RICHTER

Department of Computer Science, Natural Language Processing, Leipzig University

AND

ROELAND VAN HOUT

Centre for Language Studies, Radboud University Nijmegen

(*Received* 12 *May* 2019 – *Revised* 11 *December* 2019 – *Accepted* 11 *December* 2019)

ABSTRACT

The present approach estimates the *strength* of intensifiers in Dutch by computing their information values in a language corpus, that is, *contextual information content* (Cohen Priva, 2008; Piantadosi, Tily, & Gibson, 2011) and *Shannon Information* (Shannon & Weaver, 1948), to respectively explain the use value and the expressive value of intensifiers when they intensify a predicative adjective. Conflicting strength values help in understanding the high number of intensifiers commonly available in particular languages and the constant need for adding new ones. Our approach underlines the relevance of two measures of information content (IC) for ranking intensifiers: (i) IC in context: the more combinatorial or transitional options an intensifier has, the higher its contextual information content and consequently its use value; and (ii) IC in relation to all alternative intensifiers: the higher the surprisal value that the occurrence of an intensifier evokes, the higher its expressive value. We shall investigate the validity of these two measures by researching a large corpus of Dutch tweets and shall test whether the values of these two measures can predict the stacking order in sequences of intensifiers.

KEYWORDS: intensifiers, strength, information value, corpus linguistics

1. Introduction

This paper addresses the use and expressive values of intensifiers in Dutch based on their usage profile in a language corpus. According to many studies (e.g., Tagliamonte, 2008, 2016; Hilte, Vandekerckhove, & Daelemans, 2018;

Vandekerckhove & Vercammen, 2018), the appeal of intensifiers is not only their hyperbolic power, but also their social and emotional expressiveness. Languages often have a large and constantly changing collection of old and new intensifiers. Another salient property of intensifiers is that they can be stacked in sequences (e.g., Vandekerckhove & Vercammen, 2018). Why is it that combinations like *zo mega fucking goed* (lit., 'so mega fucking good') are fairly common, while combinations such as *mega zo fucking goed* (lit., 'mega so fucking good') are, at best, awkward? We will argue that the usage profiles of intensifiers are related to the sequences in their stacks. Intensifiers can be used in combination with adverbs, adjectives, nouns, and verbs and in different constructions. We made the plain restriction to analyse intensifiers in a straightforward, predicative context, the predicate being an adjective, such as *zij zijn echt zo fucking goed* (lit., 'they are really so fucking good') and not in an attributive context such as *het echt zo fucking goede boek* (lit., 'the really so fucking good book'), to be sure that the intensifiers are all functioning as adverbials in direct relation to an adjective. By selecting only predicates we have ensured that the intensifier applies to the adjective. In addition, the subject of the predicative construction was a third person plural pronoun, referring to living organisms (+animate). Predicative adjectives can be intensified without exception, which makes this construction ideal for our analyses. Given specific discourses and contexts, it is possible to intensify even non-gradable adjectives such as *rectangular* in order to express surprise, for instance when statements are made about persons – as it is the case in our test corpus – like *their faces are really rectangular*. We shall argue that the strength of intensifiers can be estimated by their information values. Two types of information are relevant here:

(i) *contextual information content* (=$\mathrm{IC}_{\mathrm{TRANS}}$), a form of conditional, Markov-like information, that is, a variant of conditional entropy. It is based on conditional probabilities, i.e., probabilities of transitions, and gives the amount of information that intensifiers convey within their (rightside) contexts. $\mathrm{IC}_{\mathrm{TRANS}}$ represents the contextual use value of intensifiers and can be defined as given in (1) (Cohen Priva, 2008; Piantadosi, Tily, & Gibson, 2011):

$$IC_{TRANS} = \mathbb{E}(-log_2(P(W = w|C = c_i))) \quad (1)$$

IC_{TRANS} is the expectation value of the information that a word w conveys in relation to its contexts. What counts as context is a matter of definition: contexts can be defined as n-gram co-occurrences of the target w, but also as syntactic contexts or even extra-sentential contexts (Levy, 2008).

(ii) *local* or *paradigmatic information* (=$\mathrm{IC}_{\mathrm{LOCAL}}$), i.e., *Shannon Information* (Shannon & Weaver, 1948), which refers to an expressive or surprisal value in competition with alternatives. This type is the information content of an

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

intensifier in relation to its competitors, all alternative intensifiers. The formula is given in (2):

$$IC_{LOCAL} = -log_2(P(W = w)) \quad (2)$$

IC_{LOCAL} is the well-known *Shannon Information* and part of classical entropy estimation (Shannon & Weaver, 1948), which estimates the average information content H of a variable in general. It measures the paradigmatic surprisal of a word w, independent from its contexts.

These two concepts of information are linked to concepts in Dahl (2004) (see also ten ten Buuren, van de Groep, Collin, Klatter, & de Hoop, 2018), i.e., the use value and expressive value of intensifiers: the use value IC_{TRANS} measures the usability in context, while the expressive value IC_{LOCAL} measures the paradigmatic strength of intensifiers. Both types of information make concrete what *surprisal* of appearance means, within contexts or given a set of intensifiers in the mental lexicon, and lay the groundwork for a cognitively based explanation of strength of intensifiers through the attention that new intensifiers attract. Within *surprisal theory* (Hale, 2001), it is stated that *surprisal* is equal to information and proportional to the processing difficulty of a sentence (Levy, 2008): the higher the uncertainty and the *surprisal* of a message is, the higher its information value.

Intensifiers in Dutch, in the Netherlands and in Flanders, were the subject of several studies (Foolen, Wottrich, & Zwets, 2016; ten Buuren et al., 2018; Vandekerckhove & Vercammen, 2018), the research question being if and how the frequency and modernity/recency of intensifiers might correspond, and how their properties relate to their strength. In Foolen et al. (2016) a positive correlation is postulated between modernity and strength. The constant appearance of new intensifiers is explained by the decreasing strength of existing, current intensifiers: their content is diluted when they are used too commonly and too frequently (Foolen et al., 2016). This hypothesis was not supported by ten Buuren et al. (2018). In an empirical study with pupils of a Dutch secondary school, the authors found that both the estimated frequency and modernity of intensifiers correlate in a positive way to their estimated strength, but the problem in interpreting these results is that frequency and modernity are also revealed to be positively correlated. The pupils evaluated all of the frequent intensifiers as being fairly modern. A relevant aspect of these studies is that the concepts involved are seen as gradual properties (Richter & van Hout, 2017). The approach in this paper adapts the idea of graduality in strength. When strength can be approximated by the concept of information based on probabilities, we can use corpus data to obtain strength values. That means that usage-based probabilities define strength, whereas in ten Buuren

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

et al. (2018) and Foolen et al. (2016) subjective ratings produced the strength values.

Information values may also model the establishment process of intensifiers in which semantic bleaching takes place: the intensifier's original, literal meaning is getting weaker until it is totally lost, leaving only the intensifying or amplifying function (Foolen et al., 2016) (see Sweetser, 1988, on bleaching as a process of meaning shift). For instance, the Dutch intensifier *zeer* 'very' and the German intensifier *sehr* 'very' are examples of common intensifiers that have lost their original meaning (see Dahl, 1979, on *very*). They have their roots in ninth-century Old High German and Old Saxon, i.e., *sēr* and *sēro*, respectively, meaning 'with pain, painful, sad, hard' (compare Old English *sār* 'painful'). In principle, there are no restrictions with regard to the original word class in creating new intensifiers, except that they are content words. Intensifiers can for instance be adjectives, such as *geniaal* 'ingenious' and *goed* 'good'; they can be nouns such as *kanker* 'cancer', *tyfus* 'typhus', and *moker* 'sledgehammer'; adverbs such as *super* 'super'; or verbs *fuck(ing)* 'fuck(ing)'. Bleaching of the original meaning goes hand in hand with a shift towards the adverbial class since intensifiers acquire the semantics of general (degree) adverbs.

We argue that, if intensifiers still possess the semantic properties of their original word classes, i.e., if bleaching is not completed, they tend to be positioned close to the adjective or attribute to be intensified and may not be the first element in a chain of intensifiers. As bleaching progresses, positional flexibility increases, as can be observed with established intensifiers.

According to Dahl (2004), the set of standard, established intensifiers does not exhibit a high diversity. That is to say, compared to the set of modern intensifiers, the standard set consists of a relatively small set of plain adverbs. In Dutch this set includes the intensifiers *erg* 'very', *heel* 'total', *zeer* 'very' (ten Buuren et al., 2018), and *zo* 'so'. These standard intensifiers do not have a high expressive value, but this is compensated by their high use value (Dahl, 2004): this means that they can be freely used in combinations with the word they intensify, in our case, predicative adjectives.

In the section that follows, we will argue that expressive value corresponds to the surprisal effect that intensifiers produce (described by Dahl, 2004, as *informational value*), given a set of alternative intensifiers. This assignment implies that recent, non-established intensifiers that we can classify as 'modern' produce a high amount of surprisal and thus have a high expressive value IC_{LOCAL} since they are unexpected given the higher probabilities of occurrence of the established intensifiers.

Use value corresponds to the degree of establishment that we want to relate to the different words that are intensified by the intensifier in question. The most evident hypothesis is that an established intensifier has a high use value,

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

but a low expressive value. Non-established intensifiers will have low use values and high expressive values, but we need to investigate how both measures interact in real data.

Another consequence of having strength values is the possibility to address the question of positional restrictions in stacks of intensifiers. In combinations of intensifiers, *echt* tends to occur on the leftmost position: *echt buitengewoon lekker* (lit., 'really extraordinary delicious'). In contrast, recent, non-established intensifiers such as *tyfus* seem to occur more often directly before the predicative adjective, when intensifiers are being combined (e.g., *very fucking nice* vs. *fucking very nice*). The pattern seems to be that stronger intensifiers would more likely occur near to the intensified adjective.

2. Hypotheses

The two types of information values can be used to formulate concrete hypotheses in relation to the expressive and use values of intensifiers:

H1: An established intensifier has a high use value and, consequently, a high IC_{TRANS};
H2: An expressive intensifier has a high IC_{LOCAL};
H3: Intensifiers with high use values have lower expressive values. This implies that IC_{TRANS} and IC_{LOCAL} are negatively correlated;
H4: Intensifiers basically have a free stacking order, but the more established an intensifier is, the more it tends to occur in the leftmost position.

The rationale of the last hypothesis is that an established intensifier may be helpful in interpreting a following word or phrase as another intensifier, if that word or phrase is not the predicative adjective. In this way, an established intensifier paves the way for a less established intensifier. New intensifiers are less known than established intensifiers and seem to have less positional and interpretational flexibility. They tend to occur directly in front of the predicative adjective. This implies that there is a preference for ICs to increase their IC_{TRANS} and IC_{LOCAL} values in a stack of intensifiers.

3. The corpus data

Our study is based on a Twitter corpus, as described in Grondelaers, van Hout, and van Halteren (2017). It is a sample from the large Twitter database available for Dutch. Twitter is an emblematic example of informal computer-mediated communication (CMC), with the prototypical features of digital writing (Crystal, 2001). One of the principles of CMC is to use expressive forms and/or signs to compensate for the absence of facial

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

expressions and intonation (Androutsopoulos, 2011). Intensifiers are a core category of lexical expressive markers that are used abundantly in CMC communication (Hilte et al., 2018).

A selection was made of tweets containing a full subject pronoun referring to the third person plural in combination with adjacent verb forms. Dutch has reduced pronouns with only a referential function, but the full pronouns additionally have a strong emphatic effect. Grondelaers et al. (2017) explored a large twitter corpus (TwiNL copus; Tjong Kim Sang & van den Bosch, 2013) to extract 14,658 Tweets with a full third person plural pronoun. The standard form of this pronoun is *zij* 'they', but the substandard variant *hun* 'them', in fact the object form, is increasingly taking over the subject function in spoken Dutch (Grondelaers et al., 2017). As half of the 14,658 occurrences were the substandard variant, this finding shows that CMC communication often triggers informal, spoken forms. These tweets happened to contain many intensifiers, and in selecting the tweets we were permissive in allowing all sorts of intensifiers, the decisive criterion being that the word in question was meant to increase the intensity of the adjective. We made a subcorpus of those tweets containing predicative adjectives with preceding intensifiers. It means that all the utterances contained the copula *zijn* 'be', being the third person plural verb form *zijn* 'are'. The total number of occurrences was 3692, of which 3177 had 1 intensifier (86.1%), 490 had 2 intensifiers (13.3%), and 25 had three intensifiers (0.6%), giving a grand total of 4232 intensifiers. That means that 28.9% of the selected tweets contained minimally one intensifier. This outcome convincingly indicates that we selected a context which triggers a productive usage of intensifiers. This conclusion is corroborated by the result that we counted 115 unique intensifiers. In this classification, repetitions were counted as one and the same intensifier. Orthographic variants were subsumed under their original form. Forms like *eeecht*, *zoooo*, and *wauuw* were respectively assigned to their basic forms, *echt*, *zo*, and *wauw*.

The predicative adjectives are preceded by between one and three intensifiers. We will refer to these positions as INT1, INT2, and INT3, respectively, where INT3 is the position directly preceding the adjective. The most frequent intensifiers were *echt* 'really' (2079 occurrences; 49.1%), *zo* 'so' (938; 22.2%), *fucking* 'fucking' (195; 4.6%), *super* 'super' (160; 3.8%), and *heel* 'totally' (116; 2.7%).

A considerable number of intensifiers – 56 (48.7%) of the 115 unique intensifiers – such as *tyfus* 'typhus', *irritant* 'irritating', *gruwelijk* 'horrible', *overdreven* 'overdone', *knetter* 'crackling', *hartstikke* 'very', *fake* 'fake', *boem* 'boom' and *vetmelig* 'fat', are *hapax legomena*, i.e., occurring just once. Modern ones such as *fake* 'fake' and *boem* 'boom' apparently are mixed up with old-fashioned ones like *hartstikke* 'very' (for this classification, see ten Buuren et al., 2018).

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

In Vandekerckhove and Vercammen (2018), the occurrences of 24 intensifiers were investigated in a Flemish chat corpus (2 million words). Half of these words do not occur in our corpus, showing clear differences in the use of intensifiers between Dutch in the Netherlands and Dutch in Flanders. A strong regional differentiation is found even within Flanders. The two most frequently used intensifiers there are the same as our two most frequent ones (*echt* 'really', and *zo* 'so'), although in the opposite order. The enormous productivity of the set of intensifiers is illustrated by the list of 200 different intensifiers presented in ten Buuren et al. (2018) for Dutch in the Netherlands.

4. Analysis and results

We computed IC_{TRANS} and IC_{LOCAL} for all 115 intensifiers. In 'Appendix 1' we give the IC_{TRANS} values. We based IC_{TRANS} on the probability transition matrix, independent of their position, in relation to all predicative adjectives. The IC_{LOCAL} values, again independent of their position in a stack of intensifiers, are listed in 'Appendix 2'. 'Appendix 1' contains values from 0 (meaning that there is only one unique combination between this intensifier and a following adjective) to 1.971 (*gewoon* 'plainly'). 'Appendix 2' contains values between a minimum of 1.025 (*echt* 'really'), meaning the minimal surprisal value, and 12.047, meaning a maximal surprisal value. The transition probabilities of the intensifiers *echt* 'really', and *tyfus* 'typhus' may give an idea of how IC_{TRANS} works. The latter intensifier has a low IC_{TRANS} since it combines solely with one element, the probability of that particular element being 1, and thus $IC_{TRANS} = 0$. In contrast, there are 2079 occurrences of *echt* 'really' in our corpus, and this intensifier has the highest number of co-occurring predicative adjectives, that is, 187. Consequently, the transition probabilities are small and the uncertainty is high. IC_{TRANS} should be also high: it is 0.570. This outcome is obviously lower than the outcome for *erg* 'very', which is 1.827, although this intensifier combines only with 17 elements. How can that be? In Table 1 we give the beginning of the transition vector for *echt* 'really'. Part of the transition vector of *erg* 'very', is given in Table 2.

Tables 1 and 2 both show a high transitional probability for the adjective *goed* 'good', but it is extremely high in the case of *echt* 'really': 0.42. Consequently, the IC value decreases substantially because it gives a fairly high certainty about the following context: in 42% of the occurrences it is the

TABLE 1. *Transition probabilities of echt 'really'*

	aardig	abnormal	afschuwelijk	allebei	asocial	goed
echt	0.000934	0.000467	0.000467	0.000467	0.000934	0.42

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

TABLE 2. *Transition probabilities of erg 'very'*

	aardig	actief	actueel	enthousiast	erg	goed
erg	0.02	0.02	0.02	0.06	0.02	0.22

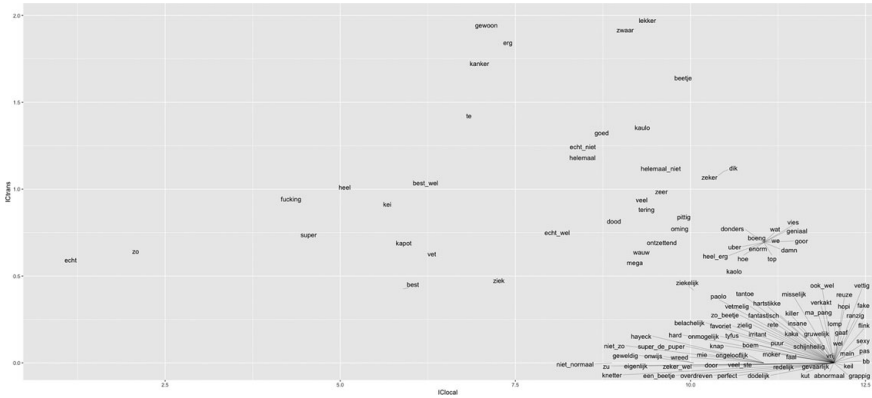

Fig. 1. Scatterplot of variables IC_{TRANS} and IC_{LOCAL}. The exceptional position of *echt* 'really' and *zo* 'so' is clearly visible. The point cloud at the bottom right consists of intensifiers with identical IC_{LOCAL} and IC_{TRANS} values. This is shown by connecting lines with identical origin.

adjective *goed* 'good'. A similar case is *zo* 'so'. There are 940 occurrences of *zo* in the corpus, but in roughly 46% of these it intensifies the adjective *goed* 'good'. Consequently, although *zo* co-occurs with several intensifiers, in almost half of its occurrences it precedes the adjective *goed* 'good', which drastically reduces its transitional uncertainty IC_{TRANS} value. In contrast, *erg* 'very' only combines with *goed* 'good' with a probability of 0.22, which increases the transitional uncertainty and thus increases the IC_{TRANS} of this intensifier. However, given our Twitter corpus, we need to consider the possibility of corpus-specific IC effects. We shall return to this point later.

How do the two IC measures correlate? The expectation is that the measures are different, as they measure two different forms of information, but they also have overlapping properties. Their correlation turns out to be negative, $r = -0.627, p = .000$, Spearman's rho even being higher, $\rho = -0.835, p = .000$, an outcome that suggests a non-linear pattern of association. The scatterplot is given in Figure 1.

On the right part of the scatterplot we see a pattern that is fairly linear, but there is a clear set of violations with intensifiers having very low IC_{LOCAL} values in combination with medium IC_{TRANS} values. The intensifiers *echt* 'really' and *zo* 'so' do not fit the overall pattern at all since these intensifiers also carry low IC_{TRANS} as low IC_{LOCAL} values.

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

Can we compare the outcomes of our two IC measures with the outcome in the two empirical studies on Dutch (ten Buuren et al., 2018; Vandekerckhove & Vercammen, 2018)? In ten Buuren et al. (2018), secondary-school children estimated the frequency and modernity of a set of intensifiers. There is an overlap of 16 intensifiers. The correlations between the IC_{TRANS} and the two estimated values are not significant. The correlations for the IC_{LOCAL} on the other hand are significant, –0.508 (p = .044) for estimated modernity, and –0.587 (p = .017) for estimated frequency. These correlations substantiate the validity of our IC_{LOCAL} measure.

For the outcomes of Vandekerckhove and Vercammen's study (2018), we observe the same pattern. Here, the frequencies of 24 intensifiers in a chat corpus for three regions in Flanders are given. There is an overlap of 12 intensifiers with the set in our study. There are, however, no significant correlations with IC_{TRANS}.

There are three (near-)significant correlations with IC_{LOCAL}: –0.690 (p = .013), West Flanders; –0.528 (p = .078), Brabant; –0.774 (p = .003), Limburg. Obviously, IC_{LOCAL} performs better, a conclusion that might be expected as our IC_{LOCAL} is also based on frequencies. It is nevertheless reassuring to see that the frequency of the occurrence of intensifiers overlaps between Dutch corpora, despite the small overlap of intensifiers involved.

Hypothesis 4 predicts preferential orders in intensifier sequences. We compared the IC values in the different positions of the intensifiers adjacent to the adjective. In Table 3 we evaluate the IC values of the three intensifier positions in terms of violating the prediction or otherwise. We compared the three positions pairwise. The percentages matched the preferential patterns predicted by our hypothesis.

Overall, we observe a strong tendency in Table 3 for the IC_{TRANS} and IC_{LOCAL} values to occur in the order predicted. The figures are more positive for IC_{LOCAL}, with 8.5% violations, than for IC_{TRANS}, with 17.0% violations. In our data, stacks of three intensifiers occur in just 25 out of 3774 tweets.

TABLE 3. *IC_{TRANS} values are predicted to decrease and IC_{LOCAL} values are predicted to increase the closer an intensifier is to the adjective; 'yes' means that the two values involved have the predicted order, 'no' means a violation*

pattern	trans			local		
INT position	yes	no	% correct	yes	no	% correct
1 versus 2	19	4	82.6%	15	8	65.2%
1 versus 3	16	7	69.6%	15	8	65.2%
2 versus 3	424	85	83.3%	476	31	93.9%
Total	469	96	83.0%	506	47	91.5%

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

TABLE 4. *Nine triplets violating decreasing IC_{TRANS} values in triplets of intensifiers*

INT1	IC_{TRANS}	INT2	IC_{TRANS}	INT3	IC_{TRANS}
best wel	1.063	heel erg	0.693	fucking	0.972
gewoon	1.971	echt	0.57	fucking	0.972
zo beetje	0	fucking	0.972	top	0.69
echt	0.57	zo	0.659	geweldig	0
echt	0.57	heel	1.036	heel erg	0.693
echt	0.57	heel	1.036	goor	0.693
echt	0.57	zo	0.659	onwijs	0
echt	0.57	zo	0.659	ziek	0.496
geniaal	0.693	helemaal	1.209	geweldig	0

TABLE 5. *Six triplets violating decreasing IC_{LOCAL} values in triplets of intensifiers*

INT1	IC_{LOCAL}	INT2	IC_{LOCAL}	INT3	IC_{LOCAL}
best wel	6.347	heel erg	11.047	fucking	4.44
gewoon	7.24	echt	1.025	fucking	4.44
zo beetje	12.047	fucking	4.44	top	11.047
bb	12.047	heel	5.189	erg	7.292
ook wel	12.047	echt	1.025	super	4.725
geniaal	1.047	helemaal	8.462	geweldig	10.047

Focusing on violations of decreasing IC_{TRANS} values within these triplets, there are 9 violating sequences, as illustrated in Table 4.

The nine triplets in Table 4 include the intensifier *echt* 'really' 6 times: this intensifier had a remarkable position in the scattergram of Figure 1. In addition, there are three intensifiers that are somehow, at the same time, a sort of mitigator: *best wel* 'best yet', *gewoon* 'plainly', and *zo beetje* 'a little bit', which all strengthen the qualification by giving it a relative perspective. For IC_{LOCAL} there are 6 triplets violating the predicted increase in their values. These triplets are displayed in Table 5. Four of them occurred in Table 4. Again, we see the occurrence of the same three mitigators mentioned in relation to Table 4. The fourth is *ook wel* 'too indeed'.

Another remarkable pattern in Table 4 is the triplet *echt heel heel erg*. The intensifier *heel* occurs twice, in fact replicating the combination *heel erg*. Replication is a pattern that frequently occurs in another form in our database, by doubling graphemes. Doublings of graphemes may strengthen the surprisal effect in the following way. If in a message *zo* is expected, but *zoo* or *zooo* occurs, the expressive value, i.e. IC_{LOCAL}, increases, when we distinguish these patterns as different. This can be seen in the frequencies in our corpus and the IC values derived from them: $IC_{LOCAL}(zo) = 2.17$, $IC_{LOCAL}(zoo) = 5.84$,

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

IC_{LOCAL}(*zooo*) = 5.78, IC_{LOCAL} (*zoooo*) = 6.12, IC_{LOCAL}(*zooooo*) = 7.14, IC_{LOCAL}(*zoooooo*) = 8.73, and IC_{LOCAL}(*zooooooo*) = 10.46. The increase in IC values facilitates combinations such as *wauw zooooooo goed*, since IC_{LOCAL}(*wauw*) = 9.46 and does not violate the principle of ascending ICs: it holds that IC_{LOCAL}(*wauw*) < IC_{LOCAL}(*zooooooo*). In this example, repetitions of identical graphemes cause a systematic increase of IC_{LOCAL}. This interpretation suggests that orthographic variants of a specific intensifier can, in contrast to the interpretation in Vandekerckhove and Vercammen (2018), be understood as intensifiers with IC_{LOCAL} values higher than the IC value of the original intensifier.

5. Discussion and conclusion

In this paper, the strength of intensifiers was determined by their information values (Hypotheses 1 and 2). The information values were based on intensifiers occurring in a Dutch Twitter corpus. The estimated information values have been confirmed by the outcomes in other studies on Dutch intensifiers (ten Buuren et al., 2018; Vandekerckhove & Vercammen, 2018), which we take as an empirical validation of our approach.

Strength of intensifiers was broken down into two information measures, i.e., IC_{TRANS} and IC_{LOCAL}, which represent the use and expressive values of intensifiers in our Twitter corpus. Both rankings of the resulting values seem to make sense. Our study confirmed our two first hypotheses: established intensifiers have a high use value, i.e., IC_{TRANS}, whereas new, expressive intensifiers have a low IC_{LOCAL} (H1 and H2). The distinction between expressive value, i.e., IC_{LOCAL}, and use value, i.e., IC_{TRANS}, seems to capture the relationship between bleaching and establishment described by Dahl (1979). The process of getting established typically means that intensifiers become real adverbs carrying only the meaning of intensification (cf. *zeer* 'very', *zo* 'so', *heel* 'wholly', *erg* 'very'). This process of establishment presupposes a high frequency of use.

Constant and frequent use and a broadening range of combinational options make the use value IC_{TRANS} increase. Conversely, they cause the expressive value IC_{LOCAL} to decrease and, consequently, both values to correlate negatively. The increase of the use value and decrease of the expressive trigger the bleaching of the intensifiers' original meaning, that is to say, both the expressive value and the use value are achievements of intensifiers of equal semantic and pragmatic relevance. The use of two types of values leads to a paradox/conflict: an intensifier combines easily with all adjectives and is therefore recognizable and transparent (and is therefore 'bleached'); an intensifier must be powerful, expressive, convincing, and therefore new.

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

There is a clear tendency in IC_{TRANS} to deliver the value '0', that represents non-informativity, for recent and surprising intensifiers. The same, but opposite, trend is evident in IC_{LOCAL}: modern and surprising intensifiers are highly informative, while established intensifiers that have undergone bleaching have a low expressive value. IC_{LOCAL} is the form of information that we would like to identify as 'strength', in a cognitive sense. It gives a formal basis to the effect of surprisal: the attention of language recipients is higher when facing a surprising intensifier than an expected one. Bleaching is a gradual process that starts with new intensifiers and only gradually takes away their original meaning. That is to say, modern intensifiers may cause surprisal and attract attention: they may unfold a high intensifying effect while still carrying a great deal of the original meaning.

A significant, medium-sized, negative correlation emerged between IC_{TRANS} and IC_{LOCAL}, as claimed in Hypothesis 3. On the other hand, the scattergram in Figure 1 shows some strong outliers that do obstruct a pure linear interpretation. We observed that *bleached* intensifiers such as *echt* 'really' have a lower than expected score because of the high share of transitions with the predicate *goed* 'good'. These outliers could be the result of a selection bias in our corpus, because our predicates are related to a specific reference: (groups of) people, as the subject of the predicate construction. This can only be tested by using other corpora and/or by widening the constructions in which intensifiers can be used. Another outlier is *zo* 'so'. Removing both outliers from the set of intensifiers does not significantly improve the correlation between the two IC measures (without *echt* 'really' and *zo* 'so': $r = -0.708, p = .000$ (Pearson), $\rho = -0.851, p = .000$ (Spearman) vs. with *echt* 'really' and *zo* 'so': $r = -0.63, p = .000$ (Pearson), $\rho = -0.84, p = .000$ (Spearman)). Given the many intensifiers, the sizes of the correlations hardly change after removing these two outliers. It is important to note as well that distinguishing the different graphemic variants of *zo* 'so' would assign the variants higher IC_{TRANS} values, pushing them to the right, non-outlying area.

We selected all words or phrases that had some intensifying function with respect to a predicative adjective. This interpretative selection procedure may produce a rather heterogeneous set of intensifiers which is illustrated by the outliers in Figure 1 such as *echt* 'really' and *zo* 'so'. In addition, we see quite different word classes, e.g., the noun *moker* 'sledgehammer', the adverb *zo* 'so', the participle *fucking* and the adjective *geniaal* 'genius'. Do we need to distinguish different classes?[1] The majority of intensifiers in the scattergram are degree modifiers. Members of this class can directly modify adjectives and thus tend to have a high IC_{LOCAL} value, but a low IC_{TRANS} value, as the correlation

[1] One reviewer pleaded for the classification of intensifiers into the three classes *degree modifiers*, *degree heads*, and *general adverbial modifiers*.

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

between the two information measures is negative. The intensifier *zo* 'so', in contrast, has both a low IC_{LOCAL} value and a low IC_{TRANS} value and as a degree head can occur to the left of degree modifier–adjective combinations. *Zo* 'so' thus combines with saturated, non-gradable, expressions. The second outlier in the scattergram, i.e., *echt* 'really', drops even more out of the scattergram cloud and seem to form its own class, as a general adverbial modifier, putting restrictions on the stacking order. On the other hand, our point of departure in Hypothesis 4 was that the stacking order is basically free. Predictions on the stacking order, based on strength, turned out to be valid, but we also observed clear violations. We refrain from calling these violations ungrammatical, but conclude that all sorts of violations are permitted, because the driving forces in using intensifiers are surprisal and unexpectedness.

This conclusion does not preclude that there is a prototypical development of intensifiers over time. New intensifiers may develop from degree modifiers to degree heads and finally to general modifiers. As a result, the position of an intensifier in our scattergram begins to shift from the lower right to the medium left into the areas of the degree head class and finally to the general modifiers class, as has happened with *echt* 'really'. Class changes may be supported by specific patterns in information values. This needs to be investigated by using more corpora, and by exploring other contexts than the predicative adjective. Given the outcomes of our usage-based approach, we provisionally conclude that intensifier classes are fuzzy.

The positive correlation between frequency and strength observed in ten Buuren et al. (2018) was confirmed in our study for the expressive value of intensifiers, that is, IC_{LOCAL}, but not for the use value, IC_{TRANS}: a surprisal effect and thus a high IC_{LOCAL} value is achieved with rare intensifiers. IC_{LOCAL} also helps to explain the high expressive value of orthographic variants of intensifiers such as *zooooooooo* or *wauuw*. These forms occur infrequently in the corpus and their surprisal effect is high, as intended by the language producer.

We hypothesized (Hypothesis 4) a preference for an increasing amount of information from left to right in combinations or stacks of intensifiers, and predicted that the most surprising and informative intensifiers directly precede the adjective. The data confirmed our hypothesis, and more convincingly so for IC_{LOCAL} than for IC_{TRANS}. The rationale of our hypothesis was that an established intensifier may be helpful in announcing another intensifier. We also observed that mitigators like *best wel* 'best yet', *gewoon* 'plainly', and *zo beetje* (lit.) 'so little bit' seemed to strengthen the qualification by giving it a relative perspective. This relativization perspective needs further investigation.

The concept of *surprisal* in information theory corresponds to the concepts of *certainty* and *uncertainty* that are integral parts of the linguistic hedges model of Zadeh (1972). Within this theory framework, *membership functions* define certainty, i.e., probabilities of memberships, for instance, the

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

probability that an entity belongs to the set of good things, to the delicious things, to the tall beings, to the young beings, etc. Probabilities of memberships can be narrowed down by a *concentration operator* that Zadeh integrated as an exponent in membership functions. Concentration operators make the probabilities of memberships smaller. It might be interesting to find out whether this concentration operator can be linked to the way we defined the strength of intensifiers within the framework of information theory.

Finally, it is important to note that the bleaching effect in modern intensifiers is not yet very advanced and at least not completed, so that in these cases the original meaning always constitutes part of the surprisal effect. Taking into account that humans tend to make predictions from contexts when they process natural language (Hale, 2001; Staub & Clifton, 2006; Levy, 2008), we pose the following principle: if an intensifier is detected in the sentence, the prediction is possible that when the next word is not the predicative adjective, it must be another (stronger) strengthening intensifier. An intensifier may even create a place for introducing new intensifiers (see Vandekerckhove & Vercammen, 2018), but, crucially, the tendencies of sequences of intensifiers discussed in this paper are not strong enough for violations to lead to ungrammaticality. This means that a sequence like *fucking zo echt goed* (lit., 'fucking so really good') or even *kanker fucking echt goed* (lit., 'cancer fucking really good') is not excluded. Such sequences are possible, though rather unusual, in current Dutch language use.

REFERENCES

Androutsopoulos, J. (2011). Language change and digital media: a review of conceptions and evidence. In T. Kristiansen Tore & N. Coupland (eds), *Standard languages and language standards in a changing Europe* (pp. 145–161). Oslo: Novus.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In N. Abner & J. Bishop (eds), *Proceedings of the 27th West Coast Conference on Formal Linguistics* (pp. 90–98). Somerville, MA: Cascadilla Proceedings Project.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Dahl, Ö. (1979). Typology of sentence negation. *Linguistics* **17**, 79–106.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. Amsterdam / Philadelphia: John Benjamins.

Foolen, A., Wottrich, V. & Zwets, M. (2016). Gruwelijk interessant: Emotieve intensiveerders in het Nederlands. Unpublished manuscript, Radboud Universiteit Nijmegen. Online https://www.ru.nl/grammarandcognition/people/vm/people/ad_foolen/publications/.

Grondelaers, S., van Hout, R. & van Halteren, H. (2017). Hun twitteren. Tweets als bron voor onderzoek naar syntactische taalvariatie. In V. De Tier, T. van de Wijngaard & A. Ghyselen (eds), *Taalvariatie en sociale media*. (pp. 65–72). Leiden: Stichting Nederlandse Dialecten.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL* (pp. 1–8). https://doi.org/10.3115/1073336.1073357

Hilte, L., Vandekerckhove, R. & Daelemans, W. (2018). Expressive markers in online teenage talk: a correlational analysis. *Nederlandse Taalkunde* **23**(3), 293–323.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177.

Piantadosi, S. T., Tily, H. & Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS* **108**(9), 3526–3529.
Richter, M. & van Hout, R. (2017). How WIE 'how' as intensifer co-occurs with other intensifiers in German sentences. In R. Loukanova & K. Liefke (eds), *Proceedings of the Workshop on Logic and Algorithms in Computational Linguistics 2017* (LACompLing2017) (pp. 133–135). Stockholm, 16–19 August 2017. Online http://su.diva-portal.org/smash/record.jsf?pid=diva2:1140018&dswid=1028.
Shannon, C. E. & Weaver, W. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423, 623–656.
Staub, A. & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: Evidence from either… or. *Journal of experimental psychology: Learning, memory, and cognition* **32**(2), 425–436.
Sweetser, E. E. (1988). Grammaticalization and semantic bleaching. *Proceedings of the fourteenth annual meeting of the Berkeley Linguistics Society* (pp. 389–405).
Tagliamonte, S. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* **12**(2), 361–394.
Tagliamonte, S. (2016). So sick or so cool? The language of youth on the internet. *Languagein Society* **45**(1), 1–32.
ten Buuren, M., van de Groep, M., Collin, S., Klatter, J. & de Hoop, H. (2018). Facking nice! Een onderzoek naar de intensiteit van intensiveerders. *Nederlandse Taalkunde* **23**, 223–250.
Tjong Kim Sang, E. & van den Bosch, A. (2013). Dealing with big data: the case of Twitter. *Computational Linguistics in the Netherlands Journal* **3**, 121–134.
Vandekerckhove, R. & Vercammen, J. (2018). De regionale en globale dynamiek van versterkers in Vlaamse jongerentaal. In T. Colleman et al. (eds), *Woorden om te bewaren: huldeboek voor Jacques Van Keymeulen* (pp. 699–712). Gent.
Zadeh, L. (1972). A fuzzy-set-theoretical interpretation of linguistic hedges. *Journal of Cybernetics* **2**, 4–34.

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

Appendix 1

The complete set of intensifiers with their IC_{TRANS}-values

Int	IC_{TRANS}	Int	IC_{TRANS}	Int	IC_{TRANS}	Int	IC_{TRANS}
abnormal	0	geweldig	0	ma pang	0	te	1.408
bb	0	gewoon	1.971	main	0	tering	0.856
beetje	1.609	goed	1.3	mega	0.556	top	0.693
belachelijk	0	goor	0.693	mie	0	tyfus	0
best	0.426	grappig	0	misselijk	0	uber	0.693
best wel	1.063	gruwelijk	0	moker	0	veel	0.963
boem	0	hard	0	niet normaal	0	veel ste	0
boeng	0.693	hartstikke	0	niet zo	0	verkakt	0
damn	0.693	hayeck	0	oming	0.746	vet	0.644
dik	1.099	heel	1.036	ongelooflijk	0	vetmelig	0
dodelijk	0	heel erg	0.693	onmogelijk	0	vettig	0
donders	0.693	helemaal	1.209	ontzettend	0.665	vies	0.693
dood	0.839	helemaal niet	1.149	onwijs	0	vrij	0
echt	0.570	hoe	0.693	ook wel	0	wat	0.693
echt niet	1.218	hopi	0	overdreven	0	wauw	0.665
echt wel	0.778	insane	0	paolo	0	we	0.693
een beetje	0	irritant	0	pas	0	wel	0
eigenlijk	0	kaka	0	perfect	0	wreed	0
enorm	0.693	kanker	1.698	pittig	0.866	zeer	0.963
erg	1.827	kaolo	0.501	puur	0	zeker	1.099
faal	0	kapot	0.665	ranzig	0	zeker wel	0
fake	0	kaulo	1.378	redelijk	0	ziek	0.496
fantastisch	0	kei	0.931	rete	0	ziekelijk	0.418
favoriet	0	killer	0	reuze	0	zielig	0
flink	0	knap	0	schijnheilig	0	zo	0.659
fucking	0.972	knetter	0	sexy	0	zo beetje	0
gaaf	0	kut	0	super	0.7	zu	0
geniaal	0.693	lekker	1.946	super de puper	0	zwaar	1.946
gevaarlijk	0	lomp	0	tantoe	0		

https://doi.org/10.1017/langcog.2019.50 Published online by Cambridge University Press

Appendix 2

The complete set of intensifiers with their IC_{LOCAL}-values

Int	IC_{LOCAL}	Int	IC_{LOCAL}	Int	IC_{LOCAL}	Int	IC_{LOCAL}
abnormal	12.047	geweldig	10.047	ma pang	12.047	te	6.762
bb	12.047	gewoon	7.240	main	12.047	tering	9.240
beetje	9.725	goed	8.588	mega	9.047	top	11.047
belachelijk	12.047	goor	11.047	mie	11.047	tyfus	12.047
best	5.897	grappig	12.047	misselijk	12.047	uber	11.047
best wel	6.347	gruwelijk	12.047	moker	12.047	veel	9.462
boem	12.047	hard	12.047	niet normaal	8.877	veel ste	12.047
boeng	11.047	hartstikke	12.047	niet zo	11.047	verkakt	12.047
damn	11.047	hayeck	11.047	oming	9.725	vet	6.214
dik	10.462	heel	5.189	ongelooflijk	11.047	vetmelig	12.047
dodelijk	12.047	heel erg	11.047	onmogelijk	12.047	vettig	12.047
donders	11.047	helemaal	8.462	ontzettend	9.462	vies	11.047
dood	9.047	helemaal niet	9.725	onwijs	10.047	vrij	12.047
echt	1.025	hoe	11.047	ook wel	12.047	wat	11.047
echt niet	8.462	hopi	12.047	overdreven	12.047	wauw	9.462
echt wel	8.240	insane	12.047	paolo	12.047	we	11.047
een beetje	12.047	irritant	12.047	pas	12.047	wel	12.047
eigenlijk	12.047	kaka	12.047	perfect	12.047	wreed	11.047
enorm	11.047	kanker	6.838	pittig	10.047	zeer	9.462
erg	7.292	kaolo	9.462	puur	12.047	zeker	10.462
faal	12.047	kapot	6.047	ranzig	12.047	zeker wel	12.047
fake	12.047	kaulo	9.462	redelijk	12.047	ziek	7.403
fantastisch	12.047	kei	5.780	rete	12.047	ziekelijk	10.047
favoriet	12.047	killer	12.047	reuze	12.047	zielig	12.047
flink	12.047	knap	12.047	schijnheilig	12.047	zo	2.174
fucking	4.440	knetter	12.047	sexy	12.047	zo beetje	12.047
gaaf	12.047	kut	12.047	super	4.425	zu	12.047
geniaal	11.047	lekker	9.240	super de puper	11.047	zwaar	9.240
gevaarlijk	12.047	lomp	12.047	tantoe	12.047		