Guidelines for Normalizing Dutch and English User Generated Content

version 1.0

LT3 Technical Report – LT3 14-01

Orphée De Clercq, Bart Desmet and Véronique Hoste

LT3 – Language and Translation Technology Team
Ghent University
orphee.declercq, bart.desmet, veronique.hoste@ugent.be

URL: http://www.lt3.ugent.be [1]

April 2014

[1]The reports of the LT3 Technical Report Series (ISSN 2032-9717) are available from http://www.lt3.ugent.be/en/publications/

Contents

1 Introduction **2**

1.1 Characteristics of user generated content 2

1.2 Two-step approach 3

1.3 Technical report 3

2 Guidelines **4**

2.1 General overview 4

2.2 In-depth explanation 5

2.2.1 MAIN COLUMNS 5

2.2.2 FLAG COLUMNS 6

2.3 Agreements 9

Chapter 1

Introduction

Before training an automatic normalization system it is crucial to manually normalize noisy data into its standard form and create a gold standard. These guidelines were developed in order to standardize the manual normalization of Dutch and English user generated content. For Dutch, these guidelines have been drawn up in close collaboration with the developers of the Chatty Corpus (Kestemont et al., 2012) and for English findings from previous studies (Baron, 2003) were included.

1.1 Characteristics of user generated content

We start by presenting some Dutch examples which clearly illustrate the main characteristics of this genre of text in Table 1.1 below.

	ORIGINAL	NORMALIZED	TRANSLATED
SMS	Oguz ! Edde me Jana gesproke ? En ze flipt lyk omdak ghsmoord heb .. !	Oh gods ! Heb je met Jana gesproken ? En ze flipt gelijk omdat ik gesmoord heb ... !	Oh god ! Did you speak to Jana ? And she's flipping because I smoked ... !
SNS	schaaaat , Je komt wel boven die Blo , je et em nii nodig wie jou laat gaan is gwn DOM :p Ilovey-ouuuu hvj	schat , Je komt wel boven die Blo , je hebt hem niet nodig wie jou laat gaan is gewoon dom :p I love you hou van je	honey, You'll get over that Blo, you don't need him whoever lets you go is just stupid :p I love you I love you
TWE	@minnebelle top ! Tis voor m'n daddy !	@minnebelle top ! Het is voor m'n daddy !	@minnebelle great ! It is for my daddy !

Table 1.1: Examples of UGC from the three social media genres representing the original utterance, its normalized version and an English translation

These examples clearly illustrate the main characteristics of UGC. Some of the more well-known problems include the omission of words or characters, e.g. the omission of the final *n* in *gesproke* (Eng: *spoke* versus *spoken*). The frequent use of abbreviations and acronyms, such as *gwn*, *hvj* (Eng: *LOL*), which are highly productive. Moreover, many utterances deviate from the standard spelling at the lexical level, such as *lyk* instead of *gelijk* (Eng: *luv* versus *love*) or by writing colloquially, e.g. *et em* instead of *hebt hem* (Eng: *you iz* vs *you are*). In UGC, emotions are also expressed by using flooding (repetition of the same character or sequence, *baaaaaaby*), emoticons (*:p*) and capitalized letters (*STUPID*).

More specific to the Dutch language is the concatenation of tokens which leads to the elimination of clitics and pronouns (*Edde* instead of *Heb je*, *khou* instead *ik hou*, *Tis* instead of *Het is*). Actually, this is also quite frequent in English UGC, e.g. *gimme, gonna, wannit*. Moreover, the influence of the English-speaking world on Belgium and the fact that it is a trilingual country often leads to various languages within a single utterance, which are often adapted to Dutch aspects (*Oguz, daddy, we are forever*).

1.2 Two-step approach

The guidelines can roughly be divided into two parts. The first part consists of the actual text normalization and comprises three steps: clearing all obvious tokenization problems, stating the different normalization operations and writing down the full normalized version. We allow four different operations: insertions, deletions, substitutions and transpositions. Examples of tokens requiring these operations are given below.

- INS: spoke (spoken), sis (sister)
- DEL: baaaaabyyyy (baby)
- SUB: iz (is), stoopid (stupid)
- TRANS: liek (like)

Insertions allow to indicate missing characters in a string. Deletions are used when characters should be deleted from a certain string. Substitutions are used when a character has been replaced with another similar one. Finally, transpositions are used when a combination of characters should be switched within one string.

The second part of the guidelines consists of flagging additional information that might be useful for further automatic processing purposes. Within each utterance the annotators were asked to indicate the end of a thought (to account for missing punctuation), regional words, foreign words and named entities. They could also flag words that are ungrammatical, stressed, part of a compound, used as interjections or words that require consecutive normalization operations.

1.3 Technical report

This technical report contains guidelines to normalize Dutch and English user generated content. The annotation and normalisation was first done in Excel and then transferred to Google Docs so as to make it easier to share and to comment on the performed normalisations. In the remainder of this report the guidelines are discussed in closer detail while giving sufficient examples (Chapter 2).

These elaborated guidelines would not have been possible to develop without the first Dutch version drawn up by Master's student Lien De Rieck.

Chapter 2

Guidelines

We distinguish a total of 16 standard columns. As mentioned in the Introduction the annotation process consists of two steps: the first six columns comprise the first step and will be referred to as the 'main columns', whereas the remaining ten columns will be called 'flag columns'. We will first present a general overview of all available columns (2.1) after which we zoom in on the relevant columns (2.2)

2.1 General overview

- MAIN COLUMNS
    1. **ID-column**: each text item receives an ID for further referral.
    2. **Original-column**: this column contains the user generated content text in its entirety.
    3. **Anomalous-column**: contains the original text split along whitespace. Each new row represents a new token. Please note that words immediately followed by punctuation or emoticons remain untouched in this column.
    4. **Tokenised-column**: is the first proper step to normalisation. If necessary, the token in the anomalous-column can be split. This action is used to separate two or more characters of different origins. *Hey!*, for instance' will be tokenised to *Hey !*. The same holds for *Jupjup!:)*, which will be converted into *Jupjup ! :)*.
    5. **Operation-column**: contains the operations needed to convert the original token into its normalised version. These operations include deletion, insertion, substitution and transposition.
    6. **Normalized-column**: consists of the normalised or standardised version of the token in the anomalous-column.
- FLAG COLUMNS
    7. **End of Thought-column**: indicates where the 'thought' should end (i.e. where we flag the end), or ends (i.e. where the actual punctuation is present).
    8. **Regional-column**: indicates whether a token is regional.
    9. **Foreign-column**: indicates whether a token is of foreign origin.
    10. **Named Entity-column**: indicates whether a token is a named entity.
    11. **Rubbish-column**: is used for tokens that have no significant meaning.

12. **Interjection-column**: indicates whether the token is an interjection.
13. **Ungrammatical-column**: indicates whether the token is ungrammatical
14. **Stressed-column**: indicates whether the token is emphasized.
15. **Compound-column**: indicates that two or more consecutive tokens are part of a compound.
16. **Multiple Normalizations-column**: indicates whether more than one operation is needed to convert the original token to its normalised version.

2.2 In-depth explanation

2.2.1 MAIN COLUMNS

An illustration of the main columns is represented in Figure 1. Since columns one, two and three are quite self-explanatory, we will start with explaining the tokenisation column.

ID	Original	Anomalous	Tokenized	Operation	Normalized
4	Ja da snappek, :p zal es kijken wak staan heb :)	Ja			Ja
		da			dat
		snappek,	snappek ,	snap~~p~~ §e§k	snap ik ,
		:p			:p
		zal			zal
		es		e##s	eens
		kijken			kijken
		wak		wa# #k	wat ik
		staan			staan
		heb			heb
		:)			:)

Figure 2.1: Dutch example illustrating the main six columns in GoogleDocs.

Tokenisation column

If necessary, the token in the anomalous-column can be split. This action is used to separate two or more characters of different origins. With 'two or more characters of different origin', we actually mean the combination of letters and figures, punctuation, or two or more words that can easily be distinguished. The following example will shed light on the matter: In *Klaatnogwelietsweten*, the spaces have been intentionally omitted, therefore, it will be tokenised as: *Klaat nog wel iets weten*. So, in this example the token *Klaat* is still written together since this actually constitutes a normalization problem. Please note that this does not hold for items that are written without white spaces in standard language, such as hour notations (e.g. *4u30*), date notations (e.g. *14/11*), abbreviations (e.g. *aug.*), etcetera.

Operation-column

We have distinguished four different operations in order to transform the original word into its normalized version. These operations comprise: substitution, insertion, deletion and transposition (cfr. Section 1.2). The method we use for annotating normalization, is based on a minimum

edit distance algorithm, the Levenshtein Distance in particular, which is a commonly used approach in spelling correction. This algorithm applies the four edit operations: deletion, insertion, substitution and transposition to calculate the difference or resemblance between two strings. This is done by measuring the LD (i.e. Levenshtein Distance) by means of these fours operations; deletion, insertion and substitution each have a cost of 1. Transpositions, however, have a cost of 2, as, in fact, one deletion and one insertion take place.

For each operation we have defined a different way of indicating it using symbols or tags:

- For insertions (INS) we use the **#** sign each time a character has to be inserted;
- For deletions (DEL) we wrap the character(s) to be removed in between tags (< **del** >< **/del** >);
- For substitutions (SUB) we wrap the character(s) to be moved in between § signs;
- For transpositions (TRANS) we wrap the characters to be transposed in between tags (< **trans** >< **/trans**>).

Please note that normalizing a word can require multiple operations, as illustrated in Figure 1, where transforming the word *snappek* to *snap ik* necessitates both a deletion and a substitution.

Normalized-column

The Normalised-column contains the fully normalized or standardized version of the Anomalous-column. E.g. *bdriege* will become *bedriegen* and the normalized version of *probeern* will be *proberen*.

2.2.2 FLAG COLUMNS

For all flag columns an index system, similar to the one applied for the Chatty Corpus annotations, will be used:

- If the normalized-column contains only one token that should be flagged as, for instance, 'interjection', a cross will be put in the interjection-column.
- If the normalized-column contains two or more tokens, a figure corresponding to the token's place in the cell is put in the flag column.
  - For instance; if the first token of a series of three is an interjection, the interjection-column will contain '1'.
  - If the second token of a series of three is an interjection, the interjection-column will contain '2'.
  - If the first and second token of a series of three tokens are interjections, the interjection-column will contain '1,2'.

End of Thought-column

Context is of utmost importance in the flagging of ends of thoughts. We have deliberately chosen to indicate the ends of thoughts, instead of sentence endings. Annotating sentence endings does not imply it is the end of thought. The author could, for example, end sentences with a full stop, and add one or more emoticons. In most cases, however, the emoticon or emoticons still refer to the previously stated sentence. Consequently, we cannot treat them as separate items. In *En weet*

ge al iets voor papa? Bel mij anders es :-) Xxx, the end of thought will be indicated on *:-)* and on *Xxx*. An example of an End of Thought that deviates from the original punctuation can be found in the following sentence: *Nonkel pascal, vergeet je ons niet morgenvroeg om 04.00 hr? ;)*. Here, the smiley will be marked End of Thought, and not the question mark.

Regional-column

Tokens will be indicated as regional if they do not occur in the Van Dale dictionary for Dutch text and for English in the Oxford English Dictionary. If tokens do, however, occur in the dictionary, but differ in meaning from the dictionary explanation, they will be regarded as regional as well. *subiet*, for example, can be found in the Van Dale dictionary and means *immediately*. In regional language, *subiet* and its variants have a different meaning, i.e. *in a while, later*. *cava, cva, etc.* are also considered as regional, even though they are of foreign origin. The reason behind this is also the change in meaning.

Regional tokens will be normalized to their stem, **uufflakke*, for instance, becomes *hoofdvlakke* and not *kopvlees*. To check the stem of regional tokens, following website will be used for Dutch: `http://www.vlaamswoordenboek.be/` and `http://www.urbandictionary.com/` for English.

Foreign-column

Tokens will be marked as foreign if they are of foreign origin and if the token's meaning was preserved in the target language. For Dutch, in most cases, the foreign tokens are of French or English origin whereas in (especially American) English many Spanish words occur. The foreign words will not be tokenized nor normalized, with the exception of flooding (i.e. the superfluous occurrence of one single character, such as in *mamaaaa*, *beeeeel!!!* and *omgggg*) , abbreviations and spelling. Dutchification of foreign words, too, will be indicated as foreign. Examples belonging to this category are *merci*, *thx*, *seriously* and the Dutchified *nicezen* as in *me myne nicezen t-shirt van avril sie xd*.

Named Entity-column

Names and surnames, brand names, towns and cities, names of services (varying from mobile phone providers *Proximus, Mobistar,...* to railway services *NMBS*) will all be flagged as named entities. Abbreviations or names referring to particular courses at school, etc. will not be considered as named entities, but as Rubbish (see next section).

Rubbish-column

Tokens flagged as rubbish comprise abbreviations of particular school courses and references or abbreviations limited to the work environment, information that has clearly not been written by the author of the message (e.g. citations) and items that cannot be understood without more context. Redundant repetitions of words, too, will be flagged in the Rubbish-column. In, for instance, *Dat meisje meisje daar*, the second occurrence of *meisje* will be fully deleted and subsequently flagged as Rubbish.

Interjection-column

The following items are marked as interjections:

- conversational initialisms and endings, with the exception of vocative use of (proper) nouns (grtz, hi, x, yo,...)
- emoticons (=p, :), :p, xp, xd,...)
- common internet abbreviations such as btw, lol, fb,...
- dialectical intensifiers such as *ze, hoor, he, wa, maat, etc* in Dutch.
- words that represent an expression, often onomatopoeia (zucht, pff, ...)
- words in between brackets
- small words such as ja/yes, nee/no without context.

Ungrammatical-column

The category of ungrammatical items is ample and comprises, amongst others, the omission of the subject, the main verb, or the combination of both subject and verb. Examples of these kinds of omissions are: **Ø Heb het vernomen.*, *'*Ø Toevallig geen tijd/ zin om iets te drinken?*. Note that the vocational use of language is not considered as ungrammatical. In many cases, other constituents of the sentence have been dropped, for instance, articles, pronouns and prepositions. E.g. **ksit in Ø cinema.* and **Is dat danØ 12 november da feestje voor u mama?*. All ellipses will be flagged as Ungrammatical in the row that follows the ellipses. Occasionally, there are too many words in one sentence, for example double negation. These, too, will be flagged as Ungrammatical.

Stressed-column

Tokens that are emphasized will be normalized to their standard form. For example, *ge-weldig* and *mét* - as opposed to *zonder*- will be normalized as *geweldig* and *met* respectively. Subsequently, both will be flagged as stressed. Various instances of flooding of words and of punctuation, such as *maamaaaaa*, which is normalized to *mama*, also belong in this category.

Compound-column

Separate tokens that are actually parts of compounds will be flagged as compounds. Note that this does not hold for Named Entities. *paard* and *rijden* in *Ik ga gn paard rijden* will both receive a Compound flag, whereas *la rocca* as in *Lynn ga ni mee & cc, noxx, la rocca & carr blijve ng over.* will not be flagged as such.

Multiple Normalizations-column

Some tokens need to be normalized in two separate steps. Instances such as *looool* and *btwww* will be marked as *Multiple Normalizations* and normalized to *laughing out loud* and *by the way*. The operations column should contain the normalization of both steps. In the case of *looool* this will be: *l#######< del> oooo < /del > o## l###* .

2.3 Agreements

Following items have been agreed upon:[1]

- Capital letters will only be preserved if they appear in the original message. When the anomalous token indicating sentence beginning contains a capital letter that is not in its right place in the normalized-column, the first letter of that sentence will be capitalized nevertheless. Following example will clarify this rather abstract description:
- Whitespaces are not insertions
- Use of regional language will be preserved *ge*, for example, will not be normalized to *je*. Verbs, too, will retain their dialectical form. *Ge wilt* will be flagged as regional but will not be standardized.
- A token can contain different substitutions if there is phonemic correspondence between the anomalous form and the standardized version. *Boejemie* will therefore be written in the operations-column as *Boejemi¡del¿e¡/del¿* and normalized to *Buscemi*.

References

Baron, Naomi S. 2003. Language of the internet. *The Stanford Handbook for Language Engineers*, pages 59–127.

Kestemont, Mike, Claudia Peersman, Benny De Decker, Guy De Pauw, Kim Luyckx, Roser Morante, Frederik Vaassen, Janneke van de Loo, and Walter Daelemans. 2012. The netlog corpus. a resource for the study of flemish dutch internet language. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may.

[1] This section should keep on growing in the next versions of this report.