# Guidelines for Dutch Normalization of User Generated Content

Claudia Matos Veliz*

November 2018

## 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.

### 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 below.

| | **Original** | **Normalized** |
|---|---|---|
| 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 ... ! |
| SNS | schaaaat , Je komt wel boven die Blo , je et em nii nodig wie jou laat gaan is gwn DOM :p Iloveyouuuu hvj | schat , Je komt wel boven die Blo , je hebt hem niet nodig wie jou laat gaan is gewoon dom <emoji> I love you hou van je |
| TWE | @minnebelle top ! Tis voor m'n daddy | <user> top ! Het is voor m'n daddy ! |

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

*Any problem or doubt regarding the guidelines or the annotation process please do not hesitate to contact me. My email address: Claudia.MatosVeliz@UGent.be

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 Text Normalization

Text normalization comprises:

- clearing all obvious tokenization problems and
- writing down the full normalized version.

If the text is to noisy and it's impossible to determine the normalized text, then write down a comment explaining the problem for that sentence.

## 2 General Overview

### 2.1 Main Columns

1. **Original Text**: Noisy sentence.
2. **Normalized Text**: Copy of the noisy sentence. The annotator must make the changes within the duplicate.
3. **Comment**:Contains the reason why the current sentence can't be annotated.
    - Foreign Language: Sometimes you can found different languages in the same text. In those case annotators can determine if the majority of the text is from a foreign language skip that text, specifying the problem in the *Comment* column.
    - Not understandable text: The annotator can't normalize the text.
    - Duplicate text: In the twitter data it is possible to find two identical or pretty similar tweets. For example the same tweet with an URL at the end, or a retweet of a previous tweet. In that case the annotator only needs to translate the original one and put a comment to the others. The tweets are sorted alphabetically, so it will be easy to identify duplicates.

- Any other comment that the annotator consider necessary.

Important: Never delete text from the file. If the annotator can not annotate certain text, they must write the problem in the *Comment* column.

## 2.2 Labels

In some cases the annotators should replace the text or emoji in the normalization for one of the labels below:

| | |
|---|---|
| <user> | For any user from the data (@minnebelle, @this_is_another_user) |
| <link> | For any url in the data. |
| <emoji> | For any emoji in the data (:P, -__-, :-DDD) |
| <hashtag> | For any hashtag in the data (#minnebelle, #this_is_another_hashtag) |

If the same label repeats (for example: *<emoji><emoji><emoji> schat , Je komt wel boven die Blo*) only one instance of that label will be necessary (*<emoji> schat , Je komt wel boven die Blo*).

## 2.3 Twitter Data

The twitter data has more particularities for the annotations. Besides the duplicates problem we mention before,the emojis were not removed from the data in order to provide a resource for a better understanding of the noisy text.

It is possible then, to find a word where some letters have been replaced by emojis (for example the *o* in *love*). In those cases the annotator must replace the emoji with the corresponding letter. The rest of the emojis have to be replaced with the <emoji> tag. Emojis of the type :P, -__-, :-DDD also must be replaced with the <emoji> tag.