Finding Syntactic Characteristics of Surinamese Dutch

Erik Tjong Kim Sang  
Meertens Institute  
erikt(at)xs4all.nl

March 26, 2015

1 Introduction

Surinamese Dutch is a variant of Dutch spoken in Suriname, a former colony of The Netherlands in the north of South America. The lexical differences between Surinamese Dutch and standard Dutch have been studied, see for example [2], but we would like to know if there are striking syntactic differences between the two language variants. In this study we will use automatic methods to compare syntactic features of two texts, one written in Surinamese Dutch and one written in standard Dutch, and describe the differences.

2 Method

We selected two novels from the Digital Library for Dutch Language (dbnl.nl): Djari/Erven by Edgar Cairo (1978, 204,000 tokens, Surinamese Dutch) and Hoe duur was de suiker? by Cynthia McLeod (1987, 118,000 tokens, standard Dutch). It would have been nice to use more data. However, Surinamese Dutch is primarily a spoken language. We do not know many other written data sources of the language.

We processed both texts with Alpino, the best available syntactic parser for Dutch [4]. The software identifies syntactic classes of words, like: dog is a noun, and generates dependency relations between words, like: dog is the subject of barks which is its syntactic head. Relations
are represented by sets with three elements: relation name, the head word and the dependent word. We evaluate two different ways of representing the head word and the dependent word by using either its lemma or its syntactic class (Part-Of-Speech). This amounts to four different dependency patterns: head-dependent is either lemma-lemma, lemma-POS, POS-lemma or POS-POS.

We will compare the texts by counting the different syntactic relations and comparing their frequencies in each text. For the comparison we use the t-test in combination with additive smoothing (add 0.5 smoothing) [1]. The t-test computes scores for pairs of related frequencies with the formula $(f_1 - f_2)/\sqrt{f_1 + f_2}$ where $f_1$ and $f_2$ are the relative frequencies of a syntactic relation in two texts. After sorting the resulting t-scores from high to low, the top of the resulting list gives an indication about what relations were more frequent in the first text than could be expected based on the second reference text.

Our automatic approach for finding dialect-specific syntactic constructions brings with it a risk of false positives and false negatives. False positives, constructions which are incorrectly suggested as dialect-specific, can originate from differences in author styles and from noise. We try to minimize the effect of these errors by inspecting the suggestions. False negatives, dialect-specific constructions which the automatic method fails to identify, could be a consequence of the language parser being unable to correctly label constructions which it has not been trained for. Presently, we have no solution for this type of error.

<table>
<thead>
<tr>
<th>t-score</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>token</th>
<th>t-score</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>token</th>
<th>t-score</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>token</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99944</td>
<td>1796</td>
<td>0</td>
<td>z’n</td>
<td>0.99600</td>
<td>249</td>
<td>0</td>
<td>Willy</td>
<td>0.99115</td>
<td>112</td>
<td>0</td>
<td>Hoor</td>
</tr>
<tr>
<td>0.99925</td>
<td>1338</td>
<td>0</td>
<td>nie</td>
<td>0.99507</td>
<td>202</td>
<td>0</td>
<td>Laila</td>
<td>0.99091</td>
<td>109</td>
<td>0</td>
<td>Schoorsteen</td>
</tr>
<tr>
<td>0.99921</td>
<td>1270</td>
<td>0</td>
<td>d'r</td>
<td>0.99502</td>
<td>200</td>
<td>0</td>
<td>...!</td>
<td>0.99065</td>
<td>106</td>
<td>0</td>
<td>Baja</td>
</tr>
<tr>
<td>0.99863</td>
<td>728</td>
<td>0</td>
<td>Bo</td>
<td>0.99444</td>
<td>179</td>
<td>0</td>
<td>dinges</td>
<td>0.99020</td>
<td>101</td>
<td>0</td>
<td>god</td>
</tr>
<tr>
<td>0.99839</td>
<td>621</td>
<td>0</td>
<td>fo</td>
<td>0.99401</td>
<td>166</td>
<td>0</td>
<td>Couplet</td>
<td>0.98969</td>
<td>96</td>
<td>0</td>
<td>niemeer</td>
</tr>
<tr>
<td>0.99830</td>
<td>586</td>
<td>0</td>
<td>em</td>
<td>0.99390</td>
<td>163</td>
<td>0</td>
<td>Aaj</td>
<td>0.98913</td>
<td>91</td>
<td>0</td>
<td>ex.</td>
</tr>
<tr>
<td>0.99778</td>
<td>450</td>
<td>0</td>
<td>Mamsi</td>
<td>0.99296</td>
<td>141</td>
<td>0</td>
<td>baja</td>
<td>0.98913</td>
<td>91</td>
<td>0</td>
<td>??</td>
</tr>
<tr>
<td>0.99714</td>
<td>349</td>
<td>0</td>
<td>Gusta</td>
<td>0.99275</td>
<td>137</td>
<td>0</td>
<td>Faader</td>
<td>0.98876</td>
<td>88</td>
<td>0</td>
<td>Weideveldt</td>
</tr>
<tr>
<td>0.99653</td>
<td>287</td>
<td>0</td>
<td>Baas</td>
<td>0.99180</td>
<td>121</td>
<td>0</td>
<td>Coola</td>
<td>0.98851</td>
<td>86</td>
<td>0</td>
<td>em.</td>
</tr>
<tr>
<td>0.99620</td>
<td>262</td>
<td>0</td>
<td>wou</td>
<td>0.99153</td>
<td>117</td>
<td>0</td>
<td>neks</td>
<td>0.98824</td>
<td>84</td>
<td>0</td>
<td>!...</td>
</tr>
</tbody>
</table>

Table 1: The 30 most salient tokens of the novel Djari/Erven when compared with Hoe duur was de suiker? by the t-test. The list is a mix between proper names, like Bo and Mamsi, common Dutch words, like z’n and wou, and words from Surinamese Dutch, like nie and fo.
Table 2: The 20 most salient dependency relations using syntactic classes for the head word and the dependent word, comparing the novel *Djari/Erven* with *Hoe duur was de suiker?* with the t-test. Patterns involving punctuation signs or words unique to one text have been omitted as well as patterns with a frequency ($f_1$) smaller than 10. The pattern *adv rhd/body comp* is associated with the sentence *waar dat ze staande loerde* (*where that she standingly peeked*).

### 3 Results

We tested the comparison method by comparing the frequencies of tokens (words plus punctuation signs) in *Djari/Erven* and *Hoe duur was de suiker?*. The 30 most salient tokens in the first text, can be found in Table 1. Words appear in this list for different reasons. Names of characters are frequently used in one book but not in the other (like *Bo* and *Mamsi*). Some common Dutch words are more commonly used by one author than the other (like *z’n* and *wou*). And finally, the list also includes words typical for the language variant of the first novel, Surinamese Dutch (*nie*, *fa*, *em*, *dinges*, *aaj*, *baja*, *nêks*, *niemeer*). This test confirms that the t-test is a useful method for extracting text-specific words.

Next, we counted the syntactic dependency relations in the two texts and compared their frequencies. We started with patterns with syntactic classes (POS) as representation of words, for example *verb has an object which is a noun*. The top 20 most salient constructions with an absolute frequency ($f_1$) of at least 10 in the Surinamese Dutch text can be found in Table 2. The names of the dependency relations and the Part-Of-Speech tags are explained in appendices A and B, respectively. Dependency patterns are not enough to get an insight in the relevant syntactic constructions. We need to inspect the sentences with a construction to check if the construction truly belongs to Surinamese Dutch. For example, a sentence which
Table 3: The 20 most salient dependency relations with lemma heads of the novel *Djari*/*Erven* when compared with *Hoe duur was de suiker?* by the t-test. Patterns involving punctuation have been omitted as well as relations with heads that did not occur in the other document. The pattern *dan dp/dp verb* corresponds with the sentence *dan kijk hoe ze wegmanoeuvreert* (then look how she leaves).

matches with the top pattern, *comp dlink/nucl noun*, is *En full speed op weg!* (And full speed ahead!). However, phrases like this example sentence are valid in standard Dutch as well so the pattern is a false positive.

We checked the sentences associated with each of the twenty patterns mentioned in Table 2. Fifteen involved patterns also occur in standard Dutch while four were uninteresting for other reasons (unrelated head/dependent words, idiomatic expression, speech error or parse error). Only for one pattern, *adv rhd/body comp*, we found an interesting example sentence: *waar dat ze staande loerde* (where that she standingly peeked). This use of phrase *where that* could be an example of Surinamese Dutch although it also irregularly appears in standard Dutch.

Next, we examined the dependency patterns involving a lemma head and a dependent part-of-speech tag. The twenty most salient patterns according to their t-score, can be found in Table 3. We examined the sentences associated with these patterns as well. Many patterns proved to be related to the start of sentence. Nineteen of the patterns were related to sentences that were also valid in standard Dutch. Only the pattern *dan dp/dp verb*, was associated with sentences that did not look like standard Dutch, for example: *dan kijk hoe ze wegmanoeuvreert* (then look how she leaves). Such an imperative sentence starting with *then* could be an example of Surinamese Dutch.
Table 4: The 20 most salient dependency relations with Part-Of-Speech heads and lemma dependents of the novel Djari/Erven when compared with Hoe duur was de suiker? by the t-test. Patterns involving punctuation have been omitted as well as patterns with heads that did not occur in the other document. The pattern comp dp/dp ga corresponds with the sentence want iemand van me familie ga kom (because someone of my family goes coming)

Tables 4 and 5 contain the top twenty syntactic relations with lemma dependents and lemma dependents and heads respectively. Again some examples of Surinamese Dutch can be found here: comp dp/dp ga: want iemand van me familie ga kom! (because someone of my family goes coming) in the first table and zeg hd/mod zo: Droomboek zeg zo, dus Vrouw Couplet ook. (Droomboek says so, so Mrs Couplet too.) ga hd/vc kom: hij heb vermoeden dat die Bo ga kom (he has suspicion that that Bo goes come) in the second table. Although there seem to be few syntactic relations that are specific to Surinamese Dutch, we are able to find some of them with the t-test.

### 4 Creating a Nederlab Case

For this particular study, tasks-specific software scripts were developed and the Alpino parser was applied to the documents which were encoded in XML. These tasks require technical knowledge. It would be nice if a comparison like in this study, could have been performed by someone without technical knowledge. The Nederlab portal aims at making this possible. Ideally a linguist could provide two texts to the portal, have them analyzed by linguistic software just as described in this paper and then be able to inspect the results.
Table 5: The 20 most salient dependency relations with lemma heads and dependents of the novel *Djari/Erven* when compared with *Hoe duur was de suiker?* by the standard t-test. Patterns involving punctuation have been omitted as well as patterns with heads that did not occur in the other document. The highlighted patterns correspond with the sentences *Droomboek zeg zo*, *dus Vrouw Couplet ook* (*Droomboek says so*; *so Mrs Couplet too*) and *hij heb vermoeden dat die Bo ga kom* (*he has suspicion that that Bo goes come*).

In order to make such a comparison possible on the Nederlab portal, the following should be arranged:

1. The comparison method (t-test or something similar) should be available on the portal as an online tool
2. In the tool it should be possible to select two texts or two document collections¹.
3. The texts or document selections should either be annotated with syntactic relations or there should be an online tool which can perform this annotation
4. The comparison tool should have the option to select the annotation level that should be compared. Different levels are interesting for the comparison, for example words and syntactic relations.
5. The comparison tool should present its analysis results sorted by t-scores. It should also be possible to download the results.

¹An alternative to starting with the comparison tool is to start with a text or collection, then select the tool and finally select a second text or collection as comparison material.
6. From the comparison results it should be possible to select the sentences associated with the different items, for example the sentences that are associated with ranked words or with ranked syntactic relations.

7. In the comparison tool, it should be possible to select, highlight and save specific parts of the result list.

In the current (March 2014) configuration of Nederlab the only available online tools involve visualization. Annotation layers have been added to all available texts in Nederlab but the layer with syntactic information used in this report has not been included because it required a lot of processing time.

5 Concluding remarks

We used an automatic method for finding syntactic differences between Surinamese Dutch and standard Dutch which employs the t-test [1]. Although the method works reasonable for discovering lexical differences (30% real differences in the top thirty of the suggestions), finding syntactic differences proved to be harder. We inspected 80 syntactic relation patterns suggested by the t-test and found five real differences between the two language variants (6%). This type of comparison is an interesting user case for the Nederlab project.

There are several ways to explain the success rate difference between the two applications. First, there are probably fewer syntactic differences between the two language variants than there are lexical variants. However, if the percentages of the differences are similar then the t-test should still perform similarly for both tasks. Second, the syntactic parser, which was trained on standard Dutch, might not notice interesting syntactic properties of the language variant because it has never encountered them before. Retraining the parser for language variants is probably too big a task so this disadvantage is hard to overcome.

A third reason for the performance difference could be the complexity of the parsing task. The frequency of a syntactic patterns is influenced by different factors, for example by word frequency when patterns with lemmas are used. We have tried to minimize this effect by examining different syntactic patterns, ignoring patterns which included lemmas unique for language variants and testing variants of the t-test. Unfortunately this did not lead to higher success rates than reported here.

From our work we draw the conclusion that the t-test is useful for finding lexical and syntactic
differences between language variants and that the syntactic difference between Surinamese
Dutch and standard Dutch is most likely smaller than the lexical difference between these
two language variants.

A  Syntactic relation names

Here are the explanations of the names of the syntactic dependency relations mentioned in
Tables 2, 3, 4 and 5. The dependency relations were defined in the project Lassy and are
used by the Dutch syntactic parser Alpino. For a complete overview of these relations, see
the Lassy Annotation manual [6], appendix A2.

- app  apposition
- body body
- det  determiner
- dp   discourse part
- ld   complement related to location or direction
- mod  modifier
- nucl nucleus
- obj1 direct object
- obj2 indirect object
- predc predicative complement
- su   subject
- tag  appendix, interjection
- vc   verbal complement

In the tables, the head type is mentioned before the relation name. Most often the head
type is head (hd) but sometimes it is different:

- cmp  complementizer
- dlink discourse link
- dp   discourse part
- hd   head
- nucl nucleus
- tag  appendix, interjection
B Part-of-Speech tags

Here is an overview of the syntactic part-of-speech tags used in Table 2, 3 and 4. These are the part-of-speech tags used by the Alpino parser, see [5] for a complete overview.

- adj adjective
- adv adverb
- comp complementizer
- det determiner
- name proper name
- noun noun
- prep preposition
- pron pronoun
- tag interjection
- verb verb

References


