SAND: Relation between the Database and Printed Maps

Erik Tjong Kim Sang
Meertens Institute
erik.tjong.kim.sang@meertens.knaw.nl

May 16, 2014

1 Introduction

SAND, the Syntactic Atlas of the Dutch Dialects, is a collection of maps of the Dutch language area (The Netherlands, Flanders and a small part of France) which show what syntactic constructions are acceptable in the dialects spoken at different locations [2]. The map data are based on interviews taken at the different locations and the results of these interviews are stored in a relational database which can be accessed via a web interface: DynaSAND [1]. The printed maps are not exactly the same as the maps shown on the website. In this report, we examine the differences and try to find out how exactly the data on the printed maps were derived from the interviews stored in the SAND database.

2 Maps which display presence and absence of constructions

The printed maps display two different types of information. Some maps display both the presence and absence of syntactic variables (positive and negative data) and others display only the presence of variables (only positive data). The difference between these two kinds of maps is important because the map data is incomplete. On maps which only display positive data, the difference is lost between a dialect without a certain construction and a dialect of which we have no information about the presence of the construction.

In this section we examine a map which contains both information about the presence and absence of a syntactic construction. For this purpose we chose the first relevant map displayed in the printed version of SAND: *interruption of the verbal cluster by a bare noun* with the test sentence *Ik weet dat Eddy morgen wil brood eten* (*I know that Eddy tomorrow wants bread eat*) [2, Volume II, page 28a, 2.3.1.1]. The printed and online maps related to this syntactic variable can be found in Figure 1. Orange/red dots and gray/blue dots indicate respectively, the presence and the absence of the variable in the local dialect. The two maps differ in two locations: Dilbeek (Flemish Brabant, Kloeke number O177p) is marked as negative on the printed map but is unmarked on the online map and Jorwerd (Friesland, B085c) is marked as negative on the printed map but is both positive and

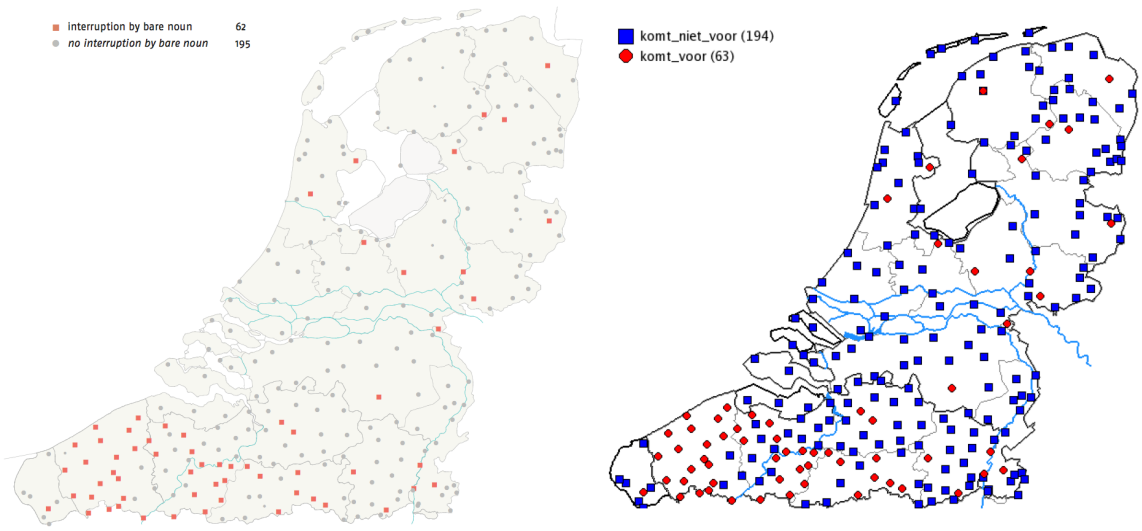

Figure 1: Maps for the variable *interruption of the verbal cluster by a bare noun* in the printed atlas (left) and the online version. Orange and red dots indicate that the interruption is allowed in the local dialect while gray and blue dots signal that the interruption is not allowed. On the left map, small gray dots represent locations for which we have no information about this syntactic variable.

negative on the online map. Let's take a look at the parts of the interviews related to this syntactic variable. First we examine Jorwerd:

1. interviewer [v=086] Ik weet dat Eddy morgen **wol** *brea* **ete**? [/v]
2. informant [a=n] Dat wol moet op t laatst van de zin. [/a]
3. interviewer [v] Ik weet dat Eddy morgen *brea* **ete wol**? [/v]
4. informant [a=j] Dat kan wel. [/a]

The interviewer presents a sentence with the bare noun *brea* interrupting the verb clause **wol ete** (2). The informant rejects the sentence and suggests placing one of the verbs (wol) at the end of the sentence (2). The interviewer repeats the sentence in the form suggested by the informant and thus without the interruption (3). The informant accepts the sentence (4). It is clear what happened here: the informant rejected [a=n] the relevant sentence [v=086] and accepted [a=j] an alternative sentence [v]. The interpretation of the printed version of the SAND seems to be correct. Now let's look at the Dilbeek interview:

1. interviewer [v=086] [/v] [a=n] Kweet dat Eddy morgen *brood* **wilt eten**. [/a]

Something strange happened here with the formatting. However from the tags between square brackets, we can derive that the informant has rejected [a=n] this example sentence

[v=086]. Again the interpretation of the printed SAND seems to be correct.

From these two examples we can infer that in the interviews example sentences are marked with [v=NNN] where NNN is a three-digit sentence number and that positive and negative answers are marked with [a=j] and [a=n] respectively. We checked this assumption for the syntactic variable examined in this section. The markers [a=j] and [a=n] only derived classifications for 246 locations where the printed map has 257 classified locations. There we take a look at some of the unclassified locations. First Midsland (Friesland, A001p):

1. interviewer [v=086] Ik weet dat Eddy morgen **wil** *brood* **eten**. [/v]
2. informant [a] Ik weet dat Eddy morgen *brood* **wil ete**. [/a]
3. interviewer *Brood* **ete wil** toch?
4. informant [a] *Brood* **ete wil**. Ja. [/a]

The informant did not repeat the example sentence and it seems that the word order is not possible in his dialect. This is the classification shown on the printed map but it is not explicitly annotated in the interview. It can only be derived by reading the interview.

Next Bakkeveen (Friesland, B127p):

1. interviewer [v=086] Ik weet dat Eddy morgen *bole* **wol ete**? [/v]
2. informant [a=j] Ja. [/a]

The informant accepted the sentence but closer inspection shows that in interviewer got the word order wrong. Therefore the judgment of the informant cannot be used and the location is left unmarked on the printed map.

Next Amsterdam (North Holland, E109p):

1. interviewer [v=086] Ik weet dat Eddy morgen **wil** *brood* **eten**. [/v]
2. informant [a=j] Dat is wel Amsterdams hoor. [/a]
3. interviewer [a] Het kan ook een woord zijn voor lunchen. [/a]
4. informant [a] Ja dat denk ik wel aan. [/a]

The example sentence is accepted by the informant but yet the location is not marked on the printed map. Perhaps the reason for this is that the example sentence was not repeated by the informant.

From the examples presented in this section, we draw the conclusion that there is no automatic method for deriving the data of the printed SAND maps from the interviews. The annotations between square brackets do not correspond completely with the printed SAND maps. In some cases interview annotations are ignored while in other cases unannotated parts of interviews have classifications on the map. As far as we have checked, the printed SAND maps contain the correct interpretations of the interviews. However these interpretations can only be made by carefully studying the conversations between the interviewers and the informants, a task which is beyond current computer software.

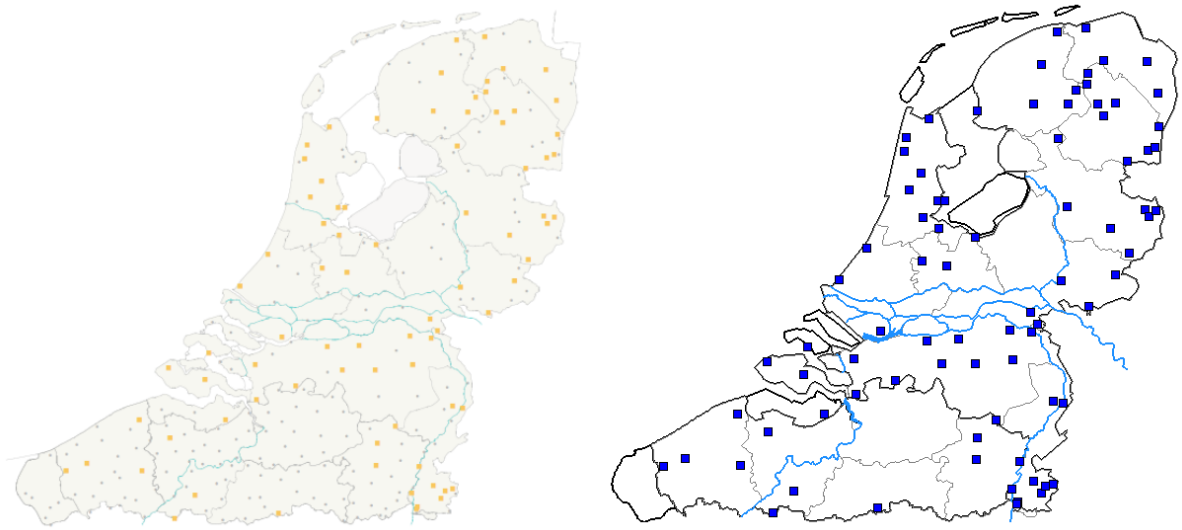

Figure 2: Maps for the variable *third construction* in the printed atlas (left) and the online version. Orange and blue dots indicate that the variable is present in the local dialect. On the left map, small gray dots represent locations for which we have no information about this syntactic variable.

3 Maps which display only the presence of constructions

Some maps in SAND only display positive cases: locations at which a certain variable is present. An example of this is the map related with the syntactic variable *third construction* which is associated with the example sentence *Wim dacht dat ik Els had geprobeerd een cadeau te geven* (*Wim thought that I Els had tried a present to give*) [2, Volume II, page 32b, 2.3.3.1]. As shown in Figure 2, the printed map for this variable (left) is exactly the same as the online map (right).

Just like for the variable we examined in the previous section, the map data cannot be derived in a straightforward way from the annotations in the interviews. For example, the relevant interview from Oosterbierum (Friesland, B046b) is:

3. interviewer [v= 531] Wim dacht dat ik els besocht hie een kadootsje te jaan? [/v]
4. informant [a=j] Dat ken. [/a]

Although both responses of the informant are marked as positive [a=j], Oosterbierum is not marked as positive on the map. The reason for this is probably that the informant did not repeat the example sentence.

For the location of Bergum (Friesland, B062p) the situation is the other way around:

1. interviewer [v=531] Wim dacht dat ik Hinke probere hie in cadeautsje te jaan? [/v]
2. informant [a] Ik leau dat it wel kan. [/a]

Although the response of the informant is not marked as positive, Bergum shows up on both maps as positive. Just like in the previous example, the example sentence was not repeated by the informant so it is unclear why exactly this location is positive on the maps.

On the two maps in Figure 2, we can only see in which locations the syntactic variable was observed in an interview. For the other locations we have no information. There are three possible explanations for an unmarked location on the map:

1. the example sentence is explicitly rejected by the informant, like in Warffum (Friesland, C029p)
2. the response of the informant to the example sentence is inconclusive, like in Schoorl (North-Holland, E017p)
3. the example sentence has not been used in the interview, like in Beetgum (Friesland, B052a)

We want to use the map data for further research and therefore it is crucial to know the difference between location without a certain syntactic variable in the local dialect (explanation 1) and locations of which we do not know if the variable is present in the local dialect (explanations 2 and 3). This difference is not encoded in the map data of the variable *third construction* and it is hard to derive the difference between explanations 1 and 2 from the interviews.

4 Maps with multi-valued syntactic variables

Most of the SAND maps contain multi-valued syntactic variables. An example of this is the map for the variable *form of the participle of the modal verb kunnen 'can'* which is associated with the example sentence *Niemand heeft dat ooit gekund* (*Nobody has that ever could*) [2, Volume II, page 37a, 2.3.6.2.2], where the final verb has twelve variants in Dutch dialects. Most interviews produced one variant but in some locations either the informant or the interviewer gave one or two alternatives.

In 31 cases variants suggested by the interviewer were rejected by the informant. The rejections are not represented on the maps. We can interpret the positive data on the maps as the preferred realizations of a syntactic variable in a dialect. By comparing these preferred choices we can still compare the data points even though we have no explicit negative information. This requires that these variables are represented as multi-valued variables rather than sets of related binary variables (the underlying binary variables would be incomplete, as explained in the previous section).

When each location contains only one alternative, this approach could work. However many locations in SAND mention more than one alternative result per variable. Spruit [3] compares the responses in two locations, Lunteren (Gelderland, F171p) and Veldhoven (North Brabant, L255p), to the test sentence *Jan herinnert zich dat verhaal wel* (*John remembers himself that story certainly*). In Lunteren, it was observed that both the word *zich* and the phrase *zijn eigen* can be used while in Veldhoven only *zich* was observed. Spruit concludes that there is a difference between the two locations but an examination of interviews reveals that this was caused by the interviewers: the one in Lunteren asked if *zich* could be replaced by *zijn eigen* while the one in Veldhoven did not. Once again the lack of an observation of a syntactic construction does not mean that the construction cannot be present in the local dialect.

5 Concluding remarks

In order to use the SAND data for automatically deriving dialect models, we need to be able to compare locations with each other. For this purpose we need to know in which locations syntactic features occur and in which they do not occur. However, explicit information about locations which do not have a particular syntactic feature, is rarely included in the SAND maps (37 of 234 maps: 16%). Negative information cannot be derived from the interviews automatically: when such information is present in the interviews, the annotation does not correspond exactly with the map data.

We recommend that the parts of the interviews related to the 20 maps with only positive information are annotated again so that explicit negative information can be added to future versions of these maps. A list of these maps can be found in the Appendix.

Most of the SAND maps (177: 76%) display syntactic variables which can take many values, like for example different verb orders. These maps rarely contain negative information. Although some negative information is present in the interviews, it is unclear how much can be extracted from them. It would be a good idea to involve some of interview parts related to these maps in the reannotation efforts. If only little negative information can be extracted, we can chose to interpret the data points as likely values for the syntactic variables at the locations and use that information for comparisons.

Appendix: SAND maps by information content

Underlined map numbers refer to map information that is absent in the SAND database.

Maps in SAND1 with only positive information (9)

43b 50b 64b 77b 79a 93a 93b 94a 95a

Maps in SAND1 with positive and negative information (13)

19a 22a 24a 26a 28a 30a 32a 87a 87b

Maps in SAND1 with multi-valued information(111)

14a 14b 15a 15b 16a 16b 17a 17b 18a 18b 20a 21a 23a 23b 25a 25b 27a 29a 29b 31a 31b 33a 33b 34a 34b 35a 35b 35c 36a 36b 38a 38b 40b 41a 42a 42b 43c 44a 46a 47a 49a 49b 50a 52a 52b 53a 53b 53c 54a 54b 55a 55b 56a 56b 56c 57a 57b 58a 58b 59a 59b 60a 61a 61b 62a 62b 63a 63b 64a 65a 65b 66a 66b 66c 68a 68b 69a 69b 70a 70b 71a 71b 72a 73a 73b 74a 74b 75a 76a 77a 78a 78b 79b 80a 80b 82a 82b 83a 84a 84b 85a 85b 86a 88a 88b 89a 89b 90a 90b 91a 91b 92a 92b 94b 95b

Maps in SAND1 with multi-valued and negative information (12)

39a 39b 40a 41b 43a 44b 45a 45b 46b 47b 48a 48b

Maps in SAND2 with only positive information (11)

32b 41b 42a 42b 48b 49a 50b 51b 53a 55a 57a

Maps in SAND2 with positive and negative information (11)

28a 28b 29a 29b 29c 30a 44a 45b 46a 46b 46c

Maps in SAND2 with multi-valued information (63)

14a 14b 15a 15b 16a 17a 17b 18a 18b 19a 20a 21a 22a 23a 24a 25a 30b 31a 31b 32a 33a 33b 34a 34b 35a 35b 36a 36b 37a 37b 38a 38b 40a 40b 41a 43a 43b 44b 44c 48a 49b 50a 51a 52a 52b 53b 54a 54b 55b 56a 56b 57b 58a 58b 59a 59b 60a 61a 62a 63a 63b 64a 64b

Maps in SAND2 with multi-valued and negative information (4)

19b 20b 45a 61b

References

[1] Sjef Barbiers. Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND), 2006. http://www.meertens.knaw.nl/sand/ Accessed 2 April 2014.

[2] Sjef Barbiers, Johan van de Auwera, Hans Bennis, Eefje Boef, Gunther Vogelaer, and Margreet van der Ham. *Syntactic Atlas of the Dutch Dialects*. Amsterdam University Press, 2008.

[3] Marco René Spruit. *Quantitative perspectives on syntactic variation in Dutch dialects.* LOT, Utrecht, The Netherlands, 2008.