Discarding noise in an automatically acquired lexicon of support verb constructions

M. Begoña Villada Moirón

Humanities Computing. University of Groningen
Postbus 716. 9700 AS Groningen
The Netherlands
villada@let.rug.nl

Abstract

We applied data-driven methods to carry out automatic acquisition of Dutch prepositional support verb constructions (SVCs) in corpora (e.g., *iets in de gaten houden* ("keep an eye on something")). This paper addresses the question whether linguistic diagnostics help to discard noise from the*nbest* lists and how to (semi-)automatically apply such linguistic diagnostics to parsed corpora. We show that some of the linguistic diagnostics proposed in Hollebrandse (1993) effectively identify SVCs and contribute a modest error rate decrease.

1. Introduction

Support verb constructions (SVCs) consist of a verb with defective semantics and a lexicalized complement that may be realized by a noun, adjective or prepositional phrase. SVCs exhibit lexical affinities between the verb and one or more lexemes inside its complement. The lexicalized complement often supplies the core meaning to the whole predicate. SVCs are located in the broad spectrum between regular verb phrases and fixed multi-word lexemes (agreeing with Sag et al. (2001)). On one hand, some SVCs participate in agreement relations and exhibit (apparent) regular syntactic structure but, on the other hand, SVCs share many idiosyncratic properties with other multi-word lexemes and idioms, for instance, limited syntactic flexibility and semantic opacity (though, the latter is not compulsory).

Corpus-based automatic acquisition methods were applied in order to compile a lexicon of Dutch support verb constructions to expand the coverage of a parser and to improve parsing accuracy. The automatically acquired *nbest* lists contain noise. This paper aims at filling in an important gap in the validation of the nbest lists proposed by statistical measures used in automatic lexical acquisition. If we can eliminate the noise from the retrieved lists in a systematic way, this will produce more reliable lexica. We describe a method to discard this noise. Our method uses some of the linguistic diagnostics proposed by Hollebrandse (1993) to distinguish regular complements of a full verb (projecting a regular verb phrase) from 'fixed' arguments of a support verb (licensing an SVC). In the remainder of this section, we present the types of noise. Section 2. summarizes the linguistic diagnostics proposed by Hollebrandse (1993). We describe a method of applying the diagnostics semi-automatically in section 3. Next, we describe the evaluation of our results. Section 5. summarizes our conclusions.

We aim at compiling a lexicon of SVCs. Preliminary experiments concentrated on automatic extraction of expressions consisting of the verb *houden* ('to hold') and a prepositional phrase. Among the higher ranked expressions,[1] some show the [verb preposition] combination HOUDEN AAN that could appear in examples like (1) and (2).

(1) Ik houd me *aan die afspraak*.
I hold myself to this agreement
'I adhere to this agreement.'

(2) Die vroeg de journalist om de man
He asked the journalist in-order-to the man
aan de praat te houden.
on the talk to hold
'He asked the journalist to keep the man talking.'

As the translation indicates, *houden aan* in (1) means 'to adhere to', something different from *houden aan* in (2) ('to keep someone hanging on'). *Houden aan de praat* constitutes part of an SVC when it appears in examples like (2) above. In this case, *houden* behaves like a support verb because the verb itself does not contribute the main semantic relation denoted by the predicate. The combination of *houden* and the PP (*aan de praat*) supplies the core meaning of the predicate. *Houden* contributes tense, aspect (progressive action), aktionsart (continuation) and causation. On the contrary, when *houden aan* means 'to adhere to', the verbal lexeme denotes meaning on its own. In addition, the preposition's object NP slot is free.

The examples (1) and (2) illustrate two types of expressions: (A) a support verb construction (eg. *(iemand) aan de praat houden* 'to keep someone hanging on' in (2)) and (B) combinations of *houden* with an ordinary prepositional complement (eg. *(zich) aan de afspraak houden* 'to adhere to the agreement' in (1)). In the second case, we do not have a fully lexicalized support verb construction but a *syntactic colligation* (Sinclair, 1966) or what Everaert (1993) calls *grammatical collocation*.

In addition to PP complements of the full verb *houden* ((B) combinations), there are other types of noise in the *nbest* list:

- locative PPs (eg. *houd onder kraan* 'hold under the tap'), temporal PPs *houd op zaterdag* 'hold on Satur-

[1] Candidate expressions were ranked with the salience test used by Kilgarrif and Tugwell (2001). Salience is an adjustment to pointwise mutual information that favors frequent candidates.

day') and directional PPs (eg. *houd naar kapel* (lit. 'hold towards the chapel').

- PPs whose head PREPOSITION introduces a required complement inside a nominal or adjectival support verb construction. For example, *houd met wensen* (lit. 'hold with wishes') whose PP may occur in the expression *rekening houden met* ('take something into account').

- other adjunct PPs that are not syntactic dependents of *houden* (eg. *houd onder auspiciën* 'hold under the auspicies'). Some of them show idiosyncratic morphosyntax (*houd tot taak* ('hold as task/aim')).

2. Linguistic diagnostics

Hollebrandse (1993) motivates a distinction between Dutch full verbs and support verbs drawing on tests that check morpho-syntactic and semantic features of the expressions. Among the diagnostics proposed by Hollebrandse (1993), we selected the following:[2]

Pronominalization If the noun phrase (NP) object inside the prepositional complement can be realized as a 'clitic' (namely *'r*, *'t*, *'m*) or the referential *er* pronoun, then the combination of verb + PP is a regular verb phrase.[3] NP pronominalization is possible with some expressions like *zich aan de wet houden* ('obey the law') in (3).

(3) Hoewel niet alle rechters gelukkig zijn met
Although not all judges lucky are with
deze wet, houden de meesten zich **er** toch aan
this law, hold the most selves there rather on
'Although not all judges are lucky with this law, most of them still obey it.'

Scrambling If the PP is *scrambled*, (i.e. an adjunct is located between the PP and its head verb) then the PP is not a fixed argument of a support verb. An adjunct occurs between the PP complement *aan de regels* and its head verb *houden* in (4).

(4) Als je je niet *aan de regels* **hier én in**
If you yourself not on the rules here and in
andere landen wilt *houden*, moet je daar de
other countries want to-hold, must you there the
consequenties van dragen.
consequences from take
'Here and in other countries, if you don't adhere to the rules, you'll have to face the consequences.'

PP **over verb** In verb final contexts, if a PP dependent (not a directional one) occurs after the verb, then the verb + PP form a regular verb phrase. The PP *aan de regels* may occur outside the verb cluster as shown in (5).

(5) Vanaf 1 januari moet de luchthaven zich houden
From 1 January must the airport itself hold
aan de regels.
on to the rules
'From January 1st, the airport must adhere to the rules.'

Coordination If a PP dependent is coordinated with a regular PP complement of the same verb, then the verb is probably a full verb. Mixed coordination of a PP complement and a fixed argument is not possible. Example (6) illustrates coordination of two fixed arguments of *houden*.

(6) Ze houden elkaar **aan de gang en in**
They hold each other on the go and in
bedwang.
control
'They keep each other in motion and in control.'

Nominalization In nominalization contexts, if the PP argument follows the nominal infinitive (its verbal head), then the combination PP VERB forms a regular verb phrase. (7) is a nominalization example of a regular VP. Note the word order change in the nominalization of an SVC in (8), where the PP precedes its nominalized head.

(7) **Je** niet **houden aan de regels** van het dualisme is
Your not hold on the rules of the dualism is
de grootst mogelijke zonde.
the biggest possible transgression
'The biggest possible transgression is to not adhere to the rules of dualism.'

(8) De leden houden zich alleen bezig met
The members hold selves only busy with
het in de gaten houden van 'verdachte personen'.
the in the holes hold of suspected people
'The members keep themselves busy by keeping an eye on 'suspects'.'

Hollebrandse (1993) adds that *NP ellipsis*, WH-movement, *heavy-NP shift* and *binding phenomena* are possible in regular verb phrases but not in SVCs. Furthermore, adjectival modification, pluralization and the use of diminutive are rather restricted inside the complements of SVCs. All diagnostics are important to determine whether a PP is part of an SVC or of a regular VP; however, we concentrate on diagnostics that can be checked automatically.

3. Applying diagnostics semi-automatically

This section reports to what extent we manage to automatize the process of checking which expressions satisfy what diagnostics, by using automatically parsed data. As input, we are given a list of expressions among the top scores in the *nbest* list.

3.1. Resources and tools

We used the Twente Nieuws Corpus (TwNC), made up of newspaper text and some television news reports (Ordelman, 2002). This corpus was already tokenized and prepared for further processing. Furthermore, the TwNC Corpus was processed by an information retrieval tool called

[2] All examples are taken from the Twente Nieuws Corpus (TwNC) `http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html`.

[3] The clitics *'r*, *'t*, *'m* correspond to the accusative feminine, neuter and masculine pronouns (English, *her*, *it* and *him*).

mg (Witten et al., 1999) to extract sentences that contain certain words or word combinations.

The Alpino parser is a wide-coverage parser for Dutch informed by a lexicalized constraint-based grammar (van der Beek et al., 2002); the grammar currently licenses a variety of syntactic constructions like subordinate clauses, (indirect) questions, (free) relative clauses, a wide range of verbal and nominal complementation and modification patterns, verbal crossing-dependency constructions, extraposition, etc. The lexicon contains approximately 47,000 lemmas. Lexical entries specify (if applicable) subcategorization frames enriched with dependency relations and some lexical restrictions. Among the parser's other components there is a highly accurate POS-tagger and a maximum entropy disambiguation module that boost the reliability of the parsed output. Currently, the POS-tagger reaches 95% per tag average accuracy while using a very large tagset (Prins and van Noord, 2004). The parser reaches about 84% per sentence average concept accuracy.[4]

dt_search is a Perl script built around XPATH, a query language to formulate queries over dependency trees encoded in XML. Bouma and Kloosterman (2002) developed this tool to support treebank queries involving constituent dependency and word order constraints.

3.2. Method

To determine what diagnostics are satisfied by the expressions in the *nbest* list, first we extract sentences that include the three lemmas inside the expression from the TwNC corpus. These sentences are collected in subcorpora and parsed with the Alpino parser. Only the best parse is returned by the parser and no error correction was performed on the parsed data. A parse (encoded in XML) is represented as a syntax tree enriched with dependency relations. Next, dt_search queries are used to determine what expressions exhibit a syntactic structure with scrambling, *pp over verb*, etc. dt_search allows us to specify head-complement dependencies, lexical restrictions, linear precedence constraints and clause type restrictions.

First, we needed to know whether the selected diagnostics efficiently identify an SVC. Therefore, in our preliminary experiments we focused on expressions with *houden* ('hold'). To this end, the sentences extracted with dt_search queries were manually checked. Two native speakers determined whether (i) the retrieved sentence effectively illustrates the diagnostic being tested and, (ii) if (i) is affirmative, whether the expression has a figurative (opaque) interpretation or a literal interpretation.

3.3. Preliminary results

Pronominalization, PP over verb and the nominalization pattern point at differences between an SVC (eg. *iemand in de gaten houden* 'keep an eye on s.o.') and a regular verb phrase (eg. *zich aan de regels houden* 'adhere to the rules'). Scrambling is useful to distinguish optional adjuncts from complements, but it does not always show a distinction between regular prepositional complements and fixed arguments in an SVC. Finally, coordination is a weak test because before judging what type of coordination an expression exhibits, one needs to know whether the PP is part of a fixed expression or not. As an illustration, Table 1 shows which diagnostics are satisfied by the expressions on the left column.

4. Evaluation

To assess whether the diagnostics help to reduce the noise in automatically extracted *nbest* lists, we selected 7 Dutch support verbs. From the *nbest* list, we extracted the 100 higher ranked expressions for each of 7 verbs. Next, we randomly collected 10% of the expressions related to each verb. Thus, we had a list of 70 expressions that were ranked among the higher scores by the salience statistic.

During automatic extraction of datasets clause boundary information was ignored. For this reason, the *nbest* list contains expressions where the verb and the PP never or almost never co-occur within the same minimal clause. Applying the diagnostics to such expressions is meaningless, thus we had to remove 6 items in the test data.

4.1. Methodology

The list of 64 expressions was given to three human judges that are Dutch native speakers. They were asked to assign a '1' if they considered the expression (part of or) a lexicalized verb phrase (SVC), a '0' if they could not think of a related lexicalized phrase and a '?' if they knew a lexicalized phrase headed by a different (support) verb but with the same PP. We allowed the third judgement because some PPs co-occur with more than one support verb denoting different *aktionsart* (eg. *op bezoek krijgen*/*hebben* 'get/have a visit'). Our gold standard list consists of those expressions assigned a '1' by at least two judges or expressions assigned a '1' and a '?'. According to the statistic all the 64 expressions are SVCs. However, according to the human judgements, 54.7% of the expressions in test data are false positives (our baseline).

We took the test data (N=64) and applied all diagnostics except *coordination*. This time, the evidence retrieved was not attested by native speakers, thus we rely on the diagnostics and our tools. Expressions that allow pronominalization, scrambling, PP over verb or show the nominalization pattern V PP are false positives. Expressions that satisfy no diagnostics or only show the nominalization pattern PP V are considered true positives.

4.2. Results

Using the human judgements as reference, the diagnostics make the wrong decisions 31.2% of the time (44 true positives, 20 false positives). This also means that the diagnostics correctly assess an item among the automatically extracted expressions as a true SVC or as noise in 70% of the cases, which is a positive outcome.

Diagnostics and human judges disagree on: (i) expressions consisting of a predicative PP (*in beroering* 'in movement'), (ii) one expression whose PP may be part of an SVC (*iemand van zijn stuk brengen* 'to surprise s.o.') or a modifier with only literal interpretation, (iii) one expression misparsed by the parser that the human judges recognized as a true SVC (*niet in de kouwe kleren gaan zitten* 'to have an

[4] Concept accuracy reflects the percentage of dependency relations within a sentence that the parser got correct.

			PP	coord		nom	
Nbest candidate expression	pron	scram	over V	PP	SVP	PP V	V PP
houd aan praat ‘keep s.o. hanging on’						∗	
houd in bedwang ‘keep s.o. in control’					∗	∗	
houd in gaten ‘keep an eye on’						∗	
houd in stand ‘keep in existence’						∗	
houd voor gek ‘make a fool of’					∗	∗	
houd oogje in zeil ‘keep a good eye on’							
houd aan afspraak ‘adhere to an agreement’	∗		∗				
houd aan regels ‘adhere to the rules’	∗	∗	∗	∗			∗
houd met wensen lit. ‘keep with wishes’			∗				
houd onder auspiciën lit. ‘hold under auspices’		∗	∗	∗			
houd van sport ‘love sport’	∗	∗		∗			

Table 1: Diagnostics evidence. pron stands for pronominalization, scram for scrambling, coord checks coordination pattern (PP or an SVP(fixed argument)), nom states nominalization pattern.

effect on’) and (iv) two directional PPs evaluated as SVCs by the diagnostics (*naar bed gaan* ‘go to bed’).

4.3. Discussion

For some expressions, no evidence was found of any of the diagnostics. This can be interpreted in two ways: either the expression satisfies none of the diagnostics or our subcorpora are not representative of the phenomena.

Our method’s success is highly dependent on parsing accuracy and also on the efficiency of the search queries. If a sentence was erroneously parsed, the retrieved evidence is likely to be wrong evidence. Good search queries require good knowledge of the grammar used by the parser. The parser has trouble in deciding the PP-attachment site. Typically, the parser favors noun attachment. Consequently, a PP part of an SVC is sometimes wrongly analyzed as a nominal post-modifier. Due to this, a query stating a head complement dependence between a given verb and the target PP will not retrieve a sentence with a misparsed post-nominal modifier PP. To avoid this, the search queries are stated more generally trying to avoid many wrong hits.

5. Conclusions

Linguistic diagnostics help to discard some sources of noise from automatically acquired lexica. For us, three tests proved most useful: pronominalization, PP over verb and the nominalization pattern. Scrambling is a good test to discard expressions that include an optional adjunct. With well-defined queries applied on parsed data, the linguistic diagnostics can automatically discard much noise from the extracted *nbest* lists. The method’s success can be further improved if a human assesses the interpretation of the retrieved evidence.

6. Acknowledgments

I am indebted to my supervisors Gertjan van Noord and Gosse Bouma for lengthy discussions about the data, their advice and technical support. I also thank Leonoor van der Beek and the three anonymous reviewers for their comments. This research was supported by the PIONIER project *Algorithms for Linguistic Processing* funded by *Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (NWO).

References

Bouma, G. and G. Kloosterman, 2002. Querying dependency treebanks in XML. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V. Las Palmas de Gran Canaria, Spain.

Everaert, M., 1993. Vaste verbindingen (in woordenboeken). *Spektator*, 3:3–27.

Hollebrandse, B., 1993. *Dutch Light Verb Constructions*. Master’s thesis, Tilburg University, the Netherlands.

Kilgarrif, A. and D. Tugwell, 2001. Word sketch: Extraction & display of significant collocations for lexicography. In *Proceedings of the 39th ACL & 10th EACL -workshop ‘Collocation: Computational Extraction, Analysis and Explotation’*. Toulouse.

Ordelman, R.J.F., 2002. Twente Nieuws Corpus (TwNC). Parlevink Language Techonology Group. University of Twente.

Prins, R. and G. van Noord, 2004. Reinforcing parser preferences through tagging. *Traitement Automatique des Langues*. Accepted for Special Issue on *Evolutions of Parsing*.

Sag, Ivan, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, 2001. Multiword expressions: a pain in the neck for NLP. LinGO Working Paper No. 2001-03.

Sinclair, J.McH., 1966. Beginning the study of lexis. In C.E. Bazell, J.C. Catford, M.A.K. Halliday, and R.H. Robins (eds.), *In memory of J.R.Firth*. Longmans, pages 410–430.

van der Beek, L., G. Bouma, J. Daciuk, T. Gaustad, R. Malouf, G. van Noord, R. Prins, and B. Villada, 2002. Algorithms for Linguistic Processing NWO PIONIER Progress Report. Groningen.

Witten, I.H., A. Moffat, and T. C. Bell, 1999. *Managing Gygabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers.