Distinguishing Prepositional Complements from Fixed Arguments

M. Begoña Villada Moirón
Alfa-Informatica. University of Groningen
P.O. Box 716. 9700 AS Groningen
The Netherlands
villada@let.rug.nl

Abstract
Statistical methods were applied to carry out automatic identification of Dutch prepositional support verb constructions in corpora. The resulting nbest list consists of expressions ranked according to the association strength inferred by the salience statistic (Kilgarrif and Tugwell, 2001). This paper addresses the question whether linguistic diagnostics help to discard noise from such automatically acquired nbest lists. We automatically applied some of the linguistic diagnostics proposed in Hollebrandse (1993) that effectively identify support verb constructions among other regular complements or adjuncts. We show that the diagnostics contribute a modest error reduction.

1 Introduction

In order to compile a lexicon of Dutch prepositional support verb constructions (SVCs), we applied statistical techniques to extract such expressions from written corpora. The resulting nbest list contains noise. If we can eliminate the noise from the automatically acquired nbest lists in a systematic way, this will produce more reliable lexica. Bearing this in mind, we investigate whether linguistic diagnostics help to identify support verb constructions in the nbest list, thus showing a distinction between regular complements (and also adjunct modifiers) and prepositional phrase (PP) arguments in support verb constructions.

In the remainder of this section, we briefly delimit what is considered a support verb construction, report how we acquired the nbest list of Dutch SVCs and describe the type of noise. Section 2 summarizes the diagnostics proposed by Hollebrandse (1993) to determine whether one candidate expression is a true prepositional SVC or not. Section 3 describes our method and preliminary results. Evaluation is reported in section 4. Section 5 summarizes our conclusions.

Support verb constructions (SVCs) consist of a verb with defective semantics and a lexicalized (fixed) argument that may be realized by a noun, an adjective or a PP. SVCs exhibit lexical affinities between the verb and one or more lexemes inside their complement. The lexicalized complement often supplies the core meaning to the whole predicate. Concerning the syntax-semantics interface, SVCs are located in the broad spectrum between regular verb phrases and fixed multi-word lexemes (agreeing with Sag et al. (2001)). Some SVCs participate in agreement relations and exhibit (apparent) regular syntactic structure but, SVCs also share many idiosyncratic properties with other multi-word lexemes and idioms, for instance, limited syntactic flexibility and often semantic opacity.

Most research on automatic acquisition of collocations and multi-word lexemes extracts candidate expressions from corpora annotated with linguistic information. Statistical tests are then used to assign a score to each candidate in the dataset. This score reflects the degree of association between the candidate composite words. One can think of this score as a measure of the lexical affinity between the composite words in a potential collocation.

In our experiments, the dataset consists of instances of the pattern [VERB PREPOSITION NOUN] (e.g. *houd aan afspraak* '(lit.) hold to agreement'). All instances were extracted from an automatically parsed corpus made up of two years of the Dutch newspapers *NRC* and *De Volkskrant*, the first one part of the Twente Nieuws Corpus (TwNC) (Ordelman, 2002). At the time of pre-processing, we did not rely on the verb-complement dependencies proposed by the Alpino parser (van der Beek et al., 2002). Thus, during dataset extraction, we tallied every verb with every PP found within a sentence. Next, the 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.

Our preliminary experiments concentrated on expressions consisting of the verb *houden* ('to hold') and a PP. Among the higher ranked expressions, some show the [V P] combination *houden aan* ('adhere to') that may 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 *aan de praat* te houden.
he asked the journalist to the man on the talk to hold
'He asked the journalist to keep the man talking.'

As the translations indicate, in (1) *houden aan* means 'to adhere to' and in (2) 'to keep someone hanging on'. *Aan de praat houden* 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 but tense, aspect (progressive action), aktionsart (continuation) and causation. The combination of *houden* and the PP (*aan de praat*) supplies the core meaning of the predicate. On the contrary, when *houden aan* means 'to adhere to', the verb denotes meaning on its own. In addition, the preposition's object NP slot is free. Examples (1) and (2) illustrate two types of expressions: (A) combinations of a full verb selecting a prepositional complement and (B) support verb constructions.

In addition to PP complements, there are other types of noise in the nbest list:

- locative PPs (e.g. *houd onder kraan* 'hold under the tap'), temporal PPs (e.g. *houd op zaterdag* 'hold on Saturday') and directional PPs (e.g. *houd naar kapel* (lit.) 'hold towards the chapel')).
- PPs whose head PREPOSITION introduces a complement in a nominal or adjectival SVC, e.g. *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* (e.g. *houd onder auspiciën* 'hold under the auspicies'). Some of them show idiosyncratic morphosyntax (*houd tot taak* 'keep to the task').

2 Linguistic diagnostics

Hollebrandse (1993) motivates a distinction between Dutch full verbs (projecting regular verb phrases) and support verbs drawing on tests that check morpho-syntactic and semantic features of the expressions. 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 diminutive are rather restricted inside the complements of SVCs. Here, we concentrate on a few diagnostics that can be checked automatically.

Pronominalization. If the noun phrase (NP) object inside the prepositional complement can be realized as a clitic (namely *'r*, *'t*, *'m* that correspond to the accusative feminine, neuter and masculine unstressed pronouns) or the referential *er* pronoun, then the combination of verb + PP is a regular verb phrase. NP pronominalization is possible with some expressions like *aan de wet houden* (3). Failure to allow pronominalization indicates that the NP inside the fixed argument is not referential and the verb and PP word combination is lexicalized.

(3) Hoewel niet alle rechters gelukkig zijn met *deze wet*, houden ze zich *er* toch aan.
Although not all judges lucky are with this law, hold they selves there rather on
'Although not all judges are lucky with this law, they still follow 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. As an example, an intervening locative PP causes scrambling in (4).

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

PP over verb. In verb final contexts, if a PP dependent occurs after the verb, then the verb + PP form a regular verb phrase. Dutch PP complements may easily be located after their verbal head in a non-finite clause or in a finite subordinate clause. In (5) the PP complement *aan de regels* follows *houden*, its lexical head. According to Haeseryn et al. (1997) PPs that constitute part of a fixed expression cannot follow the verb cluster.

(5) Vanaf 1 januari moet de luchthaven zich houden *aan de regels*.
From 1 January must the airport itself hold 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) shows the coordination of two fixed PP complements of a light verb.

(6) Ze houden elkaar *aan de gang en in bedwang*.
They hold each other on the go and in 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. In theory, any complement of a verb may appear to the left or to the right of the corresponding nominalized infinitive (Haeseryn et al., 1997). Consequently, we expect that the

nominalization of a verb taking a prepositional complement shows both patterns: PP VERB and VERB PP. The nominalization pattern VERB PP can be considered a sub-case of pp over verb because the complement PP follows its lexical head. As an example, (8) includes the nominalization of the predicate *zich aan de regels houden* and (9) the nominalization corresponding to *iets in de gaten houden.*

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

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

3 Applying diagnostics automatically

To determine to what extent the diagnostics help to identify true SVCs in the nbest list, we applied the linguistic diagnostics to 100 expressions with the verb *houden* ('hold') ranked among the top scores. For each expression, all sentences including the expression's composite lemmas [VERB PREP NOUN] were extracted from the TwNC corpus, collected into subcorpora and automatically parsed. A parse tree (encoded in XML) depicts a sentence syntactic structure enriched with dependency relations. A treebank query tool allowed us to apply the diagnostics on the parsed subcorpora, thus, retrieving evidence of PP over verb, scrambling, etc. Two native speakers assessed the evidence afterwards. Villada (2004) gives a detailed description of the procedure.

3.2 Preliminary results

Pronominalization, PP over verb and the nominalization pattern point at differences between an SVC (e.g. *iemand in de gaten houden* 'keep an eye on s.o.') and a regular verb phrase (e.g. *zich aan de regels houden* 'adhere to the rules'). Scrambling effectively distinguishes 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 one needs to know whether the PP is part of a fixed expression or not before judging what type of coordination the expression exhibits. Table 1 shows which diagnostics are satisfied by the expressions on the left column.

Nbest candidate expression	pron	scram	PP over V	coord		nom	
				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 auspicien 'hold under auspices'		*	*	*			
houd van sport 'love sport'	*	*		*			

Table 1: Diagnostic evidence for a few nbest candidates. pron stands for pronominalization, scram for scrambling, coord for coordination pattern (PP or an SVP (fixed argument)) and nom for nominalization pattern.

4 Evaluation and results

To assess whether the diagnostics help to reduce the noise, we selected 7 Dutch support verbs. From the nbest list, the 100 higher ranked expressions for each of 7 verbs were extracted and 10% of the expressions related to each verb were randomly selected. 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.

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* (e.g. *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 salience 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 and discussion

Using the human judgements as reference, the diagnostics correctly classify 44 items (27 as true positives and 17 as 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.

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. Diagnostics and human judges disagree on: (i) expressions consisting of a predicative PP (e.g. *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 effect on') and (iv) two directional PPs evaluated as SVCs by the diagnostics (*naar bed gaan* 'go to bed'). Predicative PPs are also object of disagreement between human judges. The fact that predicative PPs co-occur with a limited set of verbs makes them resemble lexicalized arguments of SVCs. Whereas predicative PPs may also occur in absolute constructions, arguments in SVCs may not. With respect to the linguistic diagnostics checked, directional PPs show a similar syntactic distribution to that of fixed arguments, however the semantics of the former is fully transparent. The diagnostics fail to detect this.

5 Conclusion

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 automatically retrieved evidence.

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

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

Hollebrandse, B. 1993. *Dutch Light Verb Constructions.* Master's Thesis, Tilburg University, The Netherlands.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and van den Toorn, M. 1997. *Algemene Nederlandse Spraakkunst*. Groningen:Wolters-Noordhoff.

Kilgarrif, A. and Tugwell, D. 2001. 'Word sketch: Extraction and Display of Significant Collocations for Lexicography' in *Proceedings of the 39th ACL & 10th EACL –workshop 'Collocation: Computational Extraction, Analysis and Explotation'*, pp. 32-38. Toulouse.

Ordelman, R. 2002. Twente Nieuws Corpus (TwNC). Parlevink Language Technology Group. Twente University. For more information see http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html

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

Villada Moirón, B. 2004. 'Discarding Noise in an Automatically Acquired Lexicon of Support Verb Constructions' in *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC) 2004. Lisbon, Portugal.To appear.