A case study on supervised classification of Swedish pseudo-coordination

Malin Ahlberg, Peter Andersson, Markus Forsberg, and Nina Tahmasebi
Språkbanken, Department of Swedish, University of Gothenburg
{firstname.lastname}@svenska.gu.se

Abstract

We present a case study on supervised classification of Swedish pseudo-coordination (SPC). The classification is attempted on the type-level with data collected from two data sets: a blog corpus and a fiction corpus. Two small experiments were designed to evaluate the feasability of this task. The first experiment explored a classifier's ability to discriminate pseudo-coordinations from ordinary verb coordinations, given a small labeled data set created during the experiment. The second experiment evaluated how well the classifier performed at detecting and ranking SPCs in a set of unlabeled verb coordinations, to investigate if it could be used as a semi-automatic discovery procedure to find new SPCs.

1 Introduction

This paper describes a case study on supervised classification of Swedish complex predicate constructions, namely pseudo-coordinations (SPCs). SPCs are light verb constructions of the form V_1 *och* V_2 'V_1 and V_2', with a semantically light V_1. An example of an SPC is *Han står och stirrar bort över havet* which could be literally translated into 'He stands and stares away over the sea', but a more correct translation would be 'He is staring away over the sea', i.e., the first verb mainly adds a progressive/durative aspect to the second verb. This example illustrates one of the reasons why SPCs, as well as constructions in general, may be worth studying from a practical language technology perspective – to improve machine translation.

We use the term 'construction' as it is used within the theoretical paradigm of construction grammar. The main tenet of construction grammar is that our grammatical knowledge is made up of a taxonomic network of constructions, i.e., pairings of form and meaning (Croft, 2001; Goldberg, 2006). Moreover, no level of grammar is considered autonomous (Fried and Östman, 2004). Constructions include all dimensions of language, form includes syntax as well as phonological aspects, and meaning includes semantics and pragmatics. Early works on construction grammar restrict the notion of constructions to form-meaning pairings with some non-predictable aspect (Goldberg, 1995), but today the concept of construction has been expanded to also include pairings with compositional meaning, which "are stored as constructions even if they are fully predictable, as long as they occur with sufficient frequency" (Goldberg, 2006). SPCs are complex constructions with a partially non-compositional meaning.

Previous work on automatic identification of Swedish constructions, e.g., Forsberg et al. (2014), focus on unsupervised classification of all constructions in a language. Forsberg et al. (2014) do this by using information-theoretic measures to rank automatically generated hybrid n-grams, where the constituents of an n-gram are either lemmas or a syntactic phrases. In this paper we are interested in a particular class of constructions, namely SPCs, where we explore the use of supervised methods that rely on available linguistic knowledge about SPCs in the classification process.

1.1 Swedish pseudo-coordination (SPC)

Pseudo-coordination is not unique to Swedish, it appears in all Scandinavian languages, as well as in other languages, such as English. If we turn our attention to pseudo-coordination in Swedish, the standard grammar reference for Swedish, Teleman et al. (1999), list five classes of SPC, based on the properties of the first verb (V_1).

1. V_1 is a position verb, e.g., *sitta* 'sit', *stå* 'stand', *ligga* 'lay'...

Kristian ***står och stirrar*** *bort över havet.*
'Kristian is staring (lit. 'stands and stares') out over the sea.'

2. V_1 is a verb of movement, e.g., *kom hit* 'come here', *åka* 'go', *går ut* 'go out', *kryper in* 'crawl inside' ...
Jag tror jag ***kryper in och sträcker ut mig*** *ett slag.*
'I think I will crawl inside and stretch myself out for a while.'

3. V_1 is a verb denoting different phases of an action, e.g., börja 'begin', fortsätta 'continue', hålla på 'keep on', sluta 'stop' ...
Folk ***håller på och tar ner*** *sina parasoller.*
'People keep on (lit. 'and') taking down their umbrellas.'

4. V_1 is a verb preceding a politeness expression, e.g., *vara (hygglig)* 'be (so kind)'.
Kan du inte ***vara hygglig och köpa*** *hem mat?*
'Could you be so kind and buy home some food?'

5. V_1 is a verb denoting the channel of communication, e.g., *skriva* 'write', *ringa* 'call', *telegrafera* 'telegraph' ...
Skriv och berätta *om dina glada upplevelser.*
'Write and tell me about your happy experiences.'

1.2 Linguistic properties of SPC

Central work related to SPCs are Teleman et al. (1999), Wiklund (2007), Kvist Darnell (2008), Blensenius (2014), and Hilpert and Koops (2008). SPCs are not as well understood as similar constructions, e.g., auxiliary constructions, which have been more extensively studied. Below you find the most prominent properties that distinguish SPCs from ordinary verb coordinations, as described in the litterature.

1. It is possible to front an object or bound adverbial of V_2 that is not compatible with V_1:
Hon satt och skrev en bok.
'She sat and wrote a book.'
⇒
Det var en bok som hon satt och skrev.
'It was a book that she sat and wrote.'

2. The order of V_1 and V_2 is fixed:
Mona satt och skrev
'Mona sat and wrote.'
⇒
?*Mona skrev och satt.*
'Mona wrote and sat'

3. Some paraphrasings are blocked:
Mona satt och sydde
'Mona sat and sewed.'
⇒
?*Mona satt och hon sydde.*
'Mona sat and she sewed'

4. *både* V_1 och V_2 'both V_1 and V_2' is blocked:
?*Mona både satt och sydde*
'Mona both sat and sewed.'

5. There are usually no or few arguments between V_1 and *och*.

6. Both verb forms have identical tense, with a few exceptions where V_1 is a modal auxiliary: *måste och handla* 'lit. must (present) and shop (infinitive)' and *vill och bada* 'lit. want (present) and bath (infinitive)'.

Other criteria are based on our own observations, or a result of discussions with colleagues. An example of what came out of these discussions is the negation test: If an SPC has a negation inserted after V_1, it also negates V_2. *Hon satt inte och skrev en bok* 'She did not (sit and) write a book'. This stands in contrast to ordinary verb coordination where the negation does not affect V_2. *Hon skrattade inte och sade ingenting* 'She did not laugh and said nothing'. Another example of what came out of these discussions was that frequency counts are very important, especially the count of the V_2 verb types; when the V_1 verb is light, it is more likely to occur with a large number of V_2 types.

Most, if not all criteria, need to be fulfilled in order for a verb coordination to qualify as a SPC. But as always when dealing with real language, between the clear cases, you find a lot of variation. One problem with using some of the above criteria is that they are all negative tests, which are known to be problematic in language classification tasks. E.g., not finding *både* V_1 *och* V_2 'both V_1 and V_2' in our data collection does not at all entail that it cannot occur, only that it has not been found in the data set.

Moreover, different SPCs seem to behave somewhat differently and the dividing line between SPCs and other complex predicates are not distinct. Both lexicalized verb constructions, such as *tycka och tänka*, 'think and reflect', and auxiliary constructions, such as *sluta och (att) spela*, 'lit. stop and (to) play', behave similarly with respect to syntactic and semantic features (Teleman et al., 1999). In fact, since 'och' and 'att' are typically pronounced in the same way, verb chains with a V_1 denoting the phase of an action are often pronounced in the same way both as SPCs and auxiliaries. Wiklund (2007) calls this group of verb chains an informal and dialectal class of SPCs.

2 Methodology

The experiments are designed for supervised classification on the type level, i.e., we do not try to decide whether a particular verb coordination in a given context is an SPC, but rather whether the verb coordination, given all its contexts, tends to function as a pseudo-coordination. For this we need a labeled data set and a suitable set of features. These features were derived from previous work and adapted to our settings. The values for each feature are based on all evidence for a verb coordination in the current data set.

Once we have trained and tested our classifier on the labeled data set, we then apply the classifier on unknown instances and evaluate the top SPC candidates according to the classifier, i.e., try to use the classifier as an SPC discovery procedure.

2.1 A random forest classifier

Using the Weka tool (Hall et al., 2009), we experimented with different types of machine learning algorithms, all with similar results. A requirement was that the classifier should be able to produce a real-valued classification to enable ranking. For no other strong reason, we ended up using a random forest classifier (Breiman, 2001). A random forest classifier consists of a combination of decision trees where features are randomly extracted to build a set of decision trees. A decision tree is a tree-structured graph where each node corresponds to a test on a feature. A path from the root to a leaf represents a classification rule.

The features are decided upon beforehand and the values for each node are learned based on training data, with the aim to best separate the positive instances from the negative instances. In our case, the instances to be classified are the verb coordinations, (V_i, V_j), that are considered positive if they are in the class SPC, and negative if they do not.

The classifier is trained and tested on labeled data from both the positive and negative class. Training and testing are performed on mutually exclusive parts of the labeled data in a stratified ten-fold cross validation. The classification results are then averaged over all ten folds.

The result according to the test data is presented in a confusion matrix with four classes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The true positive and true negative classes contain those instances that have been correctly classified. The false positive class contains all non-SPC instances that have been misclassified as SPC, and conversly, the false negative class contains all SPC instances misclassified as non-SPC.

2.2 The feature set

For each verb coordination (V_i, V_j), we derived a set of features based on the evidence in our data set. Our features were derived from Hilpert and Koops (2008), Teleman et al. (1999), Tsvetkov and Wintner (2011) (a work on classifying multiword expressions, a task similar to this one), as well as our own observations.

The features generally measured closeness and order, as well as represented negative tests, and the features were real-valued features rather than binary, i.e., a test like "is the word *både* 'both' used before the verb coordination?" was translated into "how often is the word *både* used before the verb coordination?". In particular when working with unedited text such as blogs, real-valued features can help reduce the effects of noise.

The features used by our classifier are described below.

1. **frequency** Frequency of (V_1, V_2), normalized by
   - the maximum frequency of any verb coordination
   - the average frequency of all verb coordinations
2. **closeness** How often are V_1 and V_2 separated by words other than *och* 'and'?
3. **inverse order** How often does (V_2, V_1) occur in relation to the frequency of (V_1, V_2)?

4. **inverse frequency** Frequency of (V_2, V_1), normalized using the maximum frequency of any verb coordination.

5. **inverse closeness** Similar to test 2, but for (V_2, V_1).

6. **both** How often is *både* 'both' used in conjunction to the verb coordination?

7. **between** How many words appear on average between V_1 and V_2?

8. **spread** How many different V can be found with V_1? Normalized by the maximum spread of all V_1.

9. **PMI** Pointwise mutual information as $log(p(V_1, V_2)/(p(V_1) * p(V_2)))$ where $p(V_i)$ is the relative frequency of verb V_i and $p(V_1, V_2)$ is the relative frequency of the verb coordination.

10. **not** How often does the word *inte* 'not' follow V_1: V_1 *inte och* V_2?

11. **tense** How often do V_1 and V_2 share the same tense?

12. **pos tags before** Distribution of the three most common pos-tags before the verb coordination.

13. **pos tags after** Distribution of the three most common pos-tags after the verb coordination.

Since the classification is done on the type level, it is unavoidable that we sometimes misclassify individual instances. Moreover, since the extraction of verb coordinations is currently done without any sophistication, some chains of verb coordinations can be misinterpreted, e.g., *Jag var ute och gick och hittade min bok* 'I was out walking and found my book' will probably be misclassified as SPC, since *gick och hittade* is erroneously extracted, a verb coordination that tends to be an SPC.

3 Two SPC experiments

The aim of our experiments is twofold. First, we want to know how well the known properties of SPCs can be utilized for classification, i.e., can we build a classifier that can separate known SPCs from other verb coordinations? Secondly, we want to explore if a classifier trained on labeled data can be used to detect SPCs from a set of unknown verb coordinations. We do this by labeling all unknown verb coordinations as non-SPCs and feeding them to the classifier. If the classifier judges them as SPCs, they end up in the class of false positives, with confidence scores that we can use for ranking. We then evaluate if the method can be used as a semi-automatic SPC discovery procedure by investigating the top candidates of the ranking.

The experiments are performed on two different kinds of modern Swedish data sets: a blog corpus and a fiction corpus.

3.1 The data

The blog corpus, *Bloggmix*,[1] is a collection of Swedish blog texts consisting of around 505 million tokens spanning 16 years, starting in 1998. The data has been annotated automatically using the LT tools in the Korp pipeline (Borin et al., 2012).

Since blog texts are typically informal and unedited, they contain a high degree of noise, i.e., misspellings and ungrammatical language. However, since the language of blogs typically is closer to spoken language than edited texts, and SPCs tend to be more frequent in spoken language, they contain many SPCs as well as new SPC-like constructions.

The fiction corpus, *Bonniers Romaner I&II*,[2] is some decades older and contains a more standardized language use. It consists of around 11 million tokens of Swedish fiction published between 1976 and 1981.

From each of these data sets we have extracted a training set of verb coordinations occuring at least twice in the data, manually labeled as SPC or non-SPC. The SPC instances all have V_1 listed as typical SPC-verbs by Teleman et al. (1999, §17–22), such as *sitta* 'sit' and *ringa* 'call'. In the negative training set, we collected instances of the same V_1, but used with V_2 that will force a non-SPC reading, such as *ringa och skriva* 'call and write' and *sitta och ligga* 'sit and lie down'. The negative examples also consist of verb coordinations with first verbs randomly selected from the data set. To

[1] Browsable at `http://spraakbanken.gu.se/korp/`, and downloadable (in a sentence-scrambled format) at `http://spraakbanken.gu.se/eng/resources/corpus`.

[2] Browsable at `http://spraakbanken.gu.se/korp/#?corpus=romi,romii`, and downloadable (in a sentence-scrambled format) at `http://spraakbanken.gu.se/eng/resources/corpus`.

	SPC	Non-SPC
SPC	492 (TP)	55 (FN)
Non-SPC	71 (FP)	598 (TN)

Table 1: Confusion matrix for the blog data set

Precision	Recall	Class
0.874	0.899	SPC
0.916	0.894	Non-SPC
0.897	0.897	Weighted avg

Table 2: Classification results for the blog data set.

capture slight variations, we allow a maximum of three words separating V_1 and V_2, e.g., ***sitter** i soffan **och läser*** 'sits in the sofa and reads'.

The created data sets were small, but for our explorative purposes, sufficiently large to get an idea of the feasibility of the task. For the blog texts we had 669 verb coordinations marked as non-SPC with 298 unique V_1, and 547 verb coordinations marked as SPC with 16 unique V_1 that were collected from Teleman et al. (1999). For the fiction data set we had 193 verb coordinations marked as non-SPC with 121 unique V_1, and 121 verb coordinations marked as SPC with 11 unique V_1.

3.2 SPC classification: the blog data set

In order to investigate how well the classifier performs on the blog data set, we evaluated using a stratified 10-fold cross validation on the labelled data. Tables 1 and 2 show the results. The classifier was able to correctly identify 89.9% of all SPCs and 89.4% of all non-SPCs, giving us an F1-measure of 0.897.

3.3 SPC ranking of unknowns: the blog data set

In our second experiment we tested the classifier, trained on the labeled data set, on previously unknown verb coordinations that we added as non-SPC. Table 3 shows the top ranking of verb coordinations that ended up in the false positive, i.e., the verb coordinations in the unknown set that the classifier deemed SPC. We found a few SPCs in this manner, for example, V_1 such as *åka* 'go', *stanna* 'stay, stop', *dra* '(slang) go' and *fara* '(formal) go' are all examples of V_1-verbs that occur in SPCs.

After having analyzed the ranking, we found many of the verb coordinations interesting, even though not necessarily typical SPCs. To investigate this further, we conducted a manual analysis of the results and classified each verb coordination into one of five classes, defined as follows:

1. **Class 0** No SPCs, or incorrectly extracted verb coordination, e.g., V_2 is the first word in phrasal verb.

2. **Class 1** Additive lexicalized verb coordination, e.g., *äta och dricka* 'eat and drink'.

3. **Class 2** Lexicalized SPC-like verb coordination, where V_2 is semantically more prominent than V_1, e.g., *fnysa och säga* 'snort and say'.

4. **Class 3** Verb coordination with a strong tendency to be SPC.

5. **Class 4** Support verb constructions, where V_1 is a support verb. The most common use of V_1 *och* V_2 in informal texts is actually incorrectly written, and should have been V_1 *att* V_2. E.g., *försöka och träna* 'try and (meant: to) exercise'.

Since this task is hard in the general case, we decided to only evaluate a few verb coordinations, and to do it through a consensus discussion among at least three evaluators. When in doubt, sentences that contained the verb coordination were used to support a decision. For the blog data, we evaluated in total 78 of the top-ranked verb coordinations. The majority of the verb coordinations, 53 of 78 was marked as class 0, and for the remaining classes, class 1: 5, class 2: 2, class 3: 12, and class 4: 6. With the exception of class 0, the SPC class was the largest with its 12 verb coordinations. In total, 25 of the 78 verb coordinations were of interest for further analysis.

3.4 SPC classification: the fiction data set

For the fiction data set, the cross validation results in Table 5 differ only slightly from the results on the blog data set. The classifier correctly identifies 90.6% of all SPCs and 89.1% of all non-SPCs. The F1-measure of 0.898 shows that the results are comparable to those of the blog data set. The absolute number of instances that fall into different categories differ from blogs, Table 4, but are similar in relation.

Verb pair	First verb	Conf.	#pairs	Pairs
talk and decide*	prata	1.0	169	[bestämma, ljuga, trycka,...]
go and camp	åka	1.0	195	[campa, beställa, bidra,...]
work and enjoy	jobba	1.0	146	[roa, använda, stressa, ...]
see and squeeze	se	1.0	56	[klämma, uppleva, värdera, ...]
find and try	hitta	1.0	71	[prova, leka, se, ...]
go and shower	dra	1.0	73	[duscha, kolla, fortsätta, ...]
play and crack	spela	1.0	51	[spräcka, njuta, uppträda, ...]
use and put	använda	1.0	66	[ställa, upptäcka, fungera, ...]
look and laugh	kolla	1.0	72	[skratta, klappa, fylla, ...]
eat and scoff*	äta	1.0	91	[glufsade, lyssna, babbla,...]
stay and promise	stanna	1.0	57	[lova, slappa, käka, ...]

Table 3: An extract of highly ranked verb coordinations in the blog data set. **Verb coordinations** in **bold** have a strong tendency to be SPCs, and * marks interesting verb coordinations from class 1, 2 or 4.

	SPC	Non-SPC
SPC	163 (TP)	17 (FN)
Non-SPC	21 (FP)	172 (TN)

Table 4: Confusion matrix for the fiction data set

Precision	Recall	Class
0.886	0.906	SPC
0.91	0.891	Non-SPC
0.898	0.898	Weighted avg

Table 5: Classification results for the fiction data set.

3.5 SPC ranking of unknowns: the fiction data set

Table 6 shows all verb coordinations that are classified by the algorithm as a false positive with a confidence higher than or equal to 0.6.

There is one movement SPC V_1, *åka* 'go', and one phasal SPC, V_1: *stanna* 'stay, stop'. Furthermore, we find verb coordinations that show a tendency of acting in an SPC-like way, e.g., *vända och gå* 'turn around and go'. The top candidates, *kunna* 'be able to' and *ha* 'have', are errors occuring because of faulty coordination extraction – clausal coordinations have been misinterpreted as verb coordinations. For further discussion, see section 4.

We evaluated this data set in the same manner as for the blog data, through consensus voting of at least three evaluators. Again, we only evaluated a small data set, 61 verb coordinations. We found 31 of 61 in class 0; 7 in class 1; 7 in class 2; 14 in class 3; and 1 in class 4. In total, 30 of the 61 verb coordinations were of interest for further analysis. In comparison with the same experiment on blog texts, we get 20% more, however, since we are dealing we such small sets of data, it is not possible to conclude that the difference is statistically significant.

4 Discussion

Teleman et al. (1999) list a few more SPC tests. One such important test is whether the pronounciation of the first verb is stressed, but such features are unavailable in our data sets. Neither do we take into account features that require a correct parse tree, such as object extraction (see 1.2). This test was used by Hilpert and Koops (2008) in their manual classification, but the correctness of the syntactic parses available to us was not deemed high enough to measure this correctly, especially for the unstandardized language found in blog texts. Hilpert and Koops (2008) also consider adverb placement, which is a feature approximated by feature 7, see section 2.2.

The feature **spread**, which is related to the grammaticalization of V_1, counts the number of unique V_2. Frequent SPC V_1 verbs such as *sitta* 'sit' have a high V_2 count. Interestingly, empirical evidence shows that while removing this feature gives lower results in the classification, see table 7 and 8, the corresponding classifier seemed to find more interesting SPC candidates when applied to the unknown verb coordinations. Table 9 shows the ranking of the unknown verb coordinations for the blog data set, with a corresponding F1-score of 0.847 for the classifier. That is, a five point drop in F1-score gave us the possibility to better locate up-and-coming SPCs semi-automatically. Examples of first verbs found with a confidence score of 1.0 are *googla* 'to google, googling', *ramla* 'to

verb coordination	First verb	Conf.	#coordinations	Verb coordinations
can and take	kunna	1.0	12	[ta, böra, se. . .]
have and put	ha	0.9	4	[lägga, ge, ha. . .]
go and pick up	åka	0.8	5	[hämta, hälsa, spela,. . .]
stay and buy	stanna	0.7	8	[köpa, lyssna, ta, vänta. . .]
smile and say*	le	0.7	2	[säga, verka]
say and feel	säga	0.7	8	[känna, dra, visa,. . .]
see and feel	se	0.7	4	[känna, lära, erfara,. . .]
laugh and say*	skratta	0.6	1	[säga]
live and must	leva	0.6	1	[måste]
turn around and go*	vända	0.6	1	[gå]

Table 6: Ranking of unknown verb coordinations in the fiction data set with a confidence higher than or equal to 0.6. **Verb coordinations** in **bold** have a strong tendency to be SPCs. * marks interesting verb coordinations from class 1, 2 or 4.

fall', *fara* 'to go', *resa* 'to travel', *mejla* 'email', *maila* 'email (different spelling)', *trilla* 'to fall', *varda (vart)* 'to be', *vända* 'turn', and *testa* 'test'.

The verb coordination *mejla och fråga* 'email and ask' falls into the same category as *ringa och fråga* 'call and ask' or *telegrafera och skicka* 'to telegraph and send a message'. Further down the list we find the Swedish words for *emailing*, *googling*, *commenting*, and *blogging*, i.e., new forms of communication. We also find more lexicalized verb coordinations such as: *ramla och slå (sig)* 'to fall and hurt oneself', *vända och gå* 'to turn around and go'.

Similar analysis on the fiction data set did not change the ranking substantially, probably due to it being a smaller data set, possibly because the language use is more formal and less spoken-like than in blogs. This hypothesis remains to be further investigated.

Since the data sets are small, it is important to note that our results are indicative rather than conclusive. When building the labeled data set we aimed at including well-known SPCs, as described in the reference literature, into our data. To reduce the bias of the frequency of V_1 in the data set, we added verb coordinations where V_1 both occurs in SPCs and non-SPCs. E.g., both SPCs such as *sitta och titta* 'sit and look' and non-SPCs such as *sitta och ligga* 'sit and lie down' are included in our training data to reduce this bias. We also randomly sampled verb coordinations while excluding the V_1 occuring in known SPCs. A more fair sample could be created, and will be created in future work, by sampling the negative examples according to the frequency distributions of V_1 and V_2 for the SPCs.

Precision	Recall	Class
0.853	0.797	SPC
0.843	0.888	Non-SPC
0.847	0.847	Weighted avg

Table 7: Cross validation results for the blog data set without the spread feature.

Precision	Recall	Class
0.729	0.822	SPC
0.821	0.715	Non-SPC
0.772	0.767	Weighted avg

Table 8: Cross validation results for the fiction data set without the spread feature.

5 Conclusion and future work

We presented a case study on supervised classification of Swedish pseudo-coordination. The classification results with F1 measures of 0.9 based on two separate data set indicate that it is possible to automatically separate known SPCs from other verb coordinations. When applying the classifiers on unknown verb coordinations, we found that quite a few interesting verb coordinations could be captured semi-automatically using a simple discovery procedure. However, when evaluating the result manually, it became clear that many verb coordinations had as many positive as negative SPC instances, which suggests that individual instances often cannot be estimated using general tendencies. Therefore, our next step is to explore how to do the classification on the instance-level instead of the type-level, like we do here.

Instance-level judgments will be important for the future research that we have planned, which

Verb pair	First verb	Conf.	#pairs	Pairs
go and camp	åka	1.0	120	[campa, beställa, bidra,...]
wake and see*	vakna	1.0	63	[se, leva, ligga, ...]
stay and eat	stanna	1.0	65	[käka, städa, säga,...]
pack and prepare	packa	1.0	19	[förbereda, vänta, åka, ...]
eat and listen	äta	1.0	35	[lyssna, fixa, titta, ...]
fall and break*	ramla	1.0	9	[bryta, slå, skrapa, skada, skylla, ...]
go and check	dra	1.0	47	[kolla, storhandla, gymma, ...]
nod and say*	nicka	1.0	4	[säga, se, komma, ...]
go and eat	fara	1.0	30	[käka, fixa, fika, ...]
google and find	googla	1.0	9	[titta, hitta, upptäcka, ...]
mail and ask	maila	1.0	13	[vilja, tipsa, fråga, ...]
stand and wait	stog	1.0	7	[vänta, titta, kolla ...]
comment and share*	kommentera	0	14	[dela, motivera, fråga, ...]
mail and tell	mejla	1.0	7	[berätta, säga, kolla, ...]
text message and ask	messa	0.9	6	[vilja, undra, säga, ...]
talk and decide*	prata	0.9	25	[bestämma, räkna, låtsas, ...]

Table 9: An extract of highly ranked verb coordinations in the blog data set, without the spread feature. **Verb coordinations** in **bold** have a strong tendency to be SPCs. * marks interesting verb coordinations from class 1, 2 or 4.

is to investigate how to capture constructional change of SPCs by introducing a temporal dimension to the classification. If successful in this task, we will continue by investigating if we can construct a classifier that captures constructional change in general, e.g., by trying to target constructions that in some ways are similar to SPCs. An interesting question in this context becomes: given that we do not know anything at all about the existence and/or emergence of a class of SPC, is there a way to discover them?

When adding a temporal dimension, the most interesting cases are the ambiguous ones, together with the SPC-likeness of other complex predicate constructions, which may represent an ongoing change, e.g., a grammaticalization in a continuum starting from ordinary verb coordination to auxiliary-like SPCs.

Acknowledgments

The research presented here was supported by the Swedish Research Council (through the projects *Swedish FrameNet++*, contract no 2010-6013, *Towards a knowledge-based culturomics*, dnr 2012-5738), and by the Bank of Sweden Tercentenary Foundation (grant agreement P12-0076:1) (through *A Swedish Constructicon*, dnr 2010-6013), the University of Gothenburg through its support of the Centre for Language Technology and its support of Språkbanken.

References

Kristian Blensenius. 2014. Maintaining contact with pseudoprogressive pseudocoordinations: Swedish verbal coordinations with 'sit', 'stand', and 'lie' from a spatial perspective. Ms. Dept. of Swedish, University of Gothenburg.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Leo Breiman. 2001. Random forests. In *Machine Learning*, pages 5–32.

William Croft. 2001. *Radical construction grammar : syntactic theory in typological perspective*. Oxford University Press, New York.

Markus Forsberg, Richard Johansson, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson, and Julia Prentice. 2014. From construction candidates to constructicon entries: An experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1):114–135.

Mirjam Fried and Jan-Ola Östman. 2004. *Construction grammar in a cross-language perspective*. John Benjamins Pub., Amsterdam.

Adele E. Goldberg. 1995. *Constructions : a construction grammar approach to argument structure*. Univ. of Chicago Press, Chicago.

Adele E. Goldberg. 2006. *Constructions at work : the nature of generalization in language*. Oxford Univ. Press, New York.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten.

2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Martin Hilpert and Christian Koops. 2008. A quantitative approach to the development of complex predicates. The case of Swedish Pseudo-Coordination with sitta “sit”. *Diachronica*, 25(2):242–261.

Ulrika Kvist Darnell. 2008. *Pseudosamordningar i svenska : särskilt sådana med verben sitta, ligga och stå*. Institutionen för lingvistik. Stockholms universitet, Stockholm.

Ulf Teleman, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens grammatik*. Stockholm: Norstedts.

Yulia Tsvetkov and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, pages 836–845, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna-Lena Wiklund. 2007. *The syntax of tenselessness : tense/mood/aspect-agreeing infinitivals*. Mouton de Gruyter, Berlin.