

Significant Triples: Adjective+Noun+Verb Combinations

HEIKE ZINSMEISTER AND ULRICH HEID

Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

Azenbergstraße 12, 70174 Stuttgart, Germany

{zinsmeis,heid}@ims.uni-stuttgart.de

Abstract

We investigate the identification and, to some extent, the classification of collocational word groups that consist of an adjectival modifier (A), an accusative object noun (N), and a verb (V) by means of parsing a newspaper corpus with a lexicalized probabilistic grammar.¹ Data triples are extracted from the resulting Viterbi parses and subsequently evaluated by means of a statistical association measure. The extraction results are then compared to predefined descriptive classes of ANV-triples. We also use a decision tree algorithm to classify part of the data obtained, on the basis of a small set of manually classified examples.

Much of the candidate data is lexicographically relevant: the triples include idiomatic combinations (e.g. *(sich) eine goldene Nase verdienen*, ‘get rich’, lit. ‘earn oneself a golden nose’), combinations of N+V and N+A collocations (e.g. *(eine) klare Absage erteilen*, ‘refuse resolutely’, lit. ‘give a clear refusal’), next to cases where N+V or N+A collocations are found, in combination with other (not necessarily collocational) context partners.

To extract such data from text corpora, a grammar is needed that captures verb+object relations: simple pattern matching on part-of-speech shapes is not sufficient. Statistical tools then allow to order the data in a way useful for subsequent manual selection by lexicographers.

¹This work has been carried out in the context of the *Transferbereich 32: Automatische Exzerption*, a DFG-funded project aiming at the creation of support tools for the corpus-based updating of printed dictionaries in lexicography, carried out in cooperation with the publishers Langenscheidt KG and Duden BIFAB AG.

1 Introduction

Most work on corpus-based extraction and classification of multiword lexical items so far has concentrated on collocations, i.e. on word pairs with certain properties. If larger chunks have been analysed, these were mostly multiword prepositions or adverbs (e.g. *by means of*, cf. Bouma and Villada 2002), groups of verbs, prepositions and nouns (e.g. German *zur Sprache kommen*, cf. Krenn 2000, Evert and Krenn 2001, etc.) or multiword terms.

We are interested, in this paper, in triples of open class words from general language, consisting of a verb, a noun (typically the object of the verb) and an adjective (which modifies the noun). We call these triples ‘ANV-triples’. Many ANV-triples are of lexicographic interest: some of them are as such idiomatic, others are closely related with collocations. Many of them need to be captured in dictionaries, and the tools described in this paper are meant to support lexicographers in identifying and classifying ANV-triples (however, we do not intend to provide an exhaustive fully automatic account of ANV-triples).

First, we will discuss the phenomenon and define five different classes of data (cf. section 2). Then, we will introduce a method of acquiring ANV-triples from a German text corpus (section 3), by means of parsing with a lexicalized probabilistic grammar, subsequent data collection, and sorting of the extraction results by means of a statistical association measure (the log-likelihood ratio, section 4). In section 5, we describe the clustering experiments undertaken with the raw output data, and finally, we discuss the current state of our experiments (section 6) and needs for further work.

2 The data

Collocations are binary. The British contextualist tradition (cf. Firth 1957 etc.) understands collocations as binary word groups (e.g. *proud + of*, *pay + attention*, etc.), and so does the tradition of pedagogical lexicography (cf. Hausmann (1989) and now Hausmann (2003), Runcie et al. (2002)); the latter restricts collocations only to combinations of open class words and distinguishes bases and collocates (‘Basis’ and ‘Kollokator’, in Hausmann’s terms). We follow this line of thinking, assuming that collocations are habitual combinations where the collocate can not easily be substituted, but which are not necessarily all non-compositional.

Some of the ANV-triples under analysis are combinations of two collocations with the same base (cf. Heid 1994,p.231: *allgemeine Gültigkeit haben*: (*allgemein + Gültigkeit*) + (*Gültigkeit + haben*), ‘be generally valid’ (lit. ‘general validity have’)). On a descriptive level, we distinguish five different types of ANV-triples. The five types are not exclusive. There is a gradient change from one to the other.

- i. **A+N+V lexically fixed:** idiomatic phrases like *sich einen schönen Lenz machen*, ‘take it easy’ (lit. ‘oneself a nice spring make’) which are typically non-compositional in meaning.
- ii. **A+N lexically fixed; V compositional:** idiomatic or collocative adjective+noun combinations that occur with and without a verb; the A+N combination may be terminologically fixed (as in *absolute Mehrheit + erreichen* (‘win an absolute majority’)), or it may be a general language expression in itself, as in *schwarze Zahlen schreiben* (‘be profitable’: ‘black numbers write’), where *schwarze Zahlen* is an idiomatic way to express the notion of profitability.
- iii. **N+V lexically fixed; A compositional:** combinations in which a random adjective occurs with a noun+verb collocation like *einen neuen Haftbefehl erlassen* ‘issue a new warrant’ in which *Haftbefehl erlassen* is a collocation which allows to be modified.
- iv. **Combinations of collocations: N+V lexically fixed; A+N lexically fixed; same N:** a combination of two collocations, i.e. of type (ii) and type (iii), as in *ein biblisches Alter erreichen*, ‘reach

a grand old age' (lit. 'a biblical age reach') in which the adjective+noun collocation *biblisches Alter* 'grand old age' interacts with the semantically compositional noun verb collocation *ein Alter (von n Jahren) erreichen*, 'reach an age (of n years)'.

- v. **Trivial combination:** non-collocative (i.e. completely compositional and non-habitual) combinations of adjective, noun, and verb as in *neue Politik fordern* 'demand new politics'.

We thereby ignore the fact that the cooccurrence of an adjectival modifier and its modifyee, the noun, is never completely at random but restricted by semantic properties of both elements. The same holds for the cooccurrence of verbs and their accusative objects. 'Random' in our sense means that there is no idiomatic interpretation or habitual use of the combination.

3 Corpus-based acquisition

Our goal is to identify significant triples of adjectives, nouns, and verbs, and to subsequently classify them into the five different classes described above, in section 2. The basis of this undertaking is the collection of frequency data from a corpus. For this task, we employ linguistic preprocessing by means of a fully-fledged probabilistic grammar that encodes predicate argument structures and provides full sentence parses (see Schulte im Walde et al. (2001) for a general overview on the grammar model and its use for the extraction of lexical information). The grammar allows us to identify grammatical structures independently from the linear order of the elements. This is relevant especially for a language like German that allows for a relatively free word order; for illustration, see e.g. example (1) below. Complex data such as the combination of a verb, its accusative object and an adjectival modifier of the latter, cannot be collected by means of shallow parsing methods or bag-of-words approaches satisfactorily. All the more, since the combination includes three parameters that are realized by open word classes.

The probabilistic grammar is based on a manually established context-free grammar with feature constraint annotations such as the specification of the subcategorization frame at all levels of a verbal category. The rule probabilities were learned in unsupervised training on a newspaper corpus of approx. 25 million words by a probabilistic parser (Schmid 2000). The grammar is lexicalized in the course of training, which means that each rule is multiplied by all potential lexical heads. Lexical heads are the lemmas of the syntactic heads in terminal phrases which are then propagated to non-terminal structures. Lexicalization allows the grammar to learn lexical cooccurrences, i.e. head-head relations between mother nodes and their non-head daughter nodes, e.g. the relation between the verbal head of a clause and the nominal head of its accusative object.

The trained grammar model allows to directly read off estimated frequencies of pairs of lexical heads (see Schulte im Walde (2003) for an overview of the lexical information that is encoded in the model itself). For triples of lexical heads, this is not manageable, at least not if the lexical heads belong to open class categories.² We therefore reparsed the corpus with the trained grammar by using the Viterbi option of the parser that determines the most probable parse for each sentence (see e.g. Manning and Schütze 1999, 396ff.). The Viterbi parses were then stored for subsequent extraction of the frequency data.

The example (1), is taken from the newspaper *Frankfurter Rundschau*, 1992/93, illustrates the word order problems encountered in the data. It includes the idiomatic expression *rote Zahlen*

²Information about closed class items may be integrated into the grammar categories, which gives indirect access to triple information, e.g. in the case of prepositional objects: the preposition lemma is then added to the category name, like *PP.in* for PPs headed by *in*, which leaves the head feature of the PP open for the embedded nominal. This allows the grammar to learn the head-head cooccurrence of a verb and the nominal head of its prepositional object, which is then directly extractable from the grammar model. But the grammar model does not provide more complex lexical dependencies.

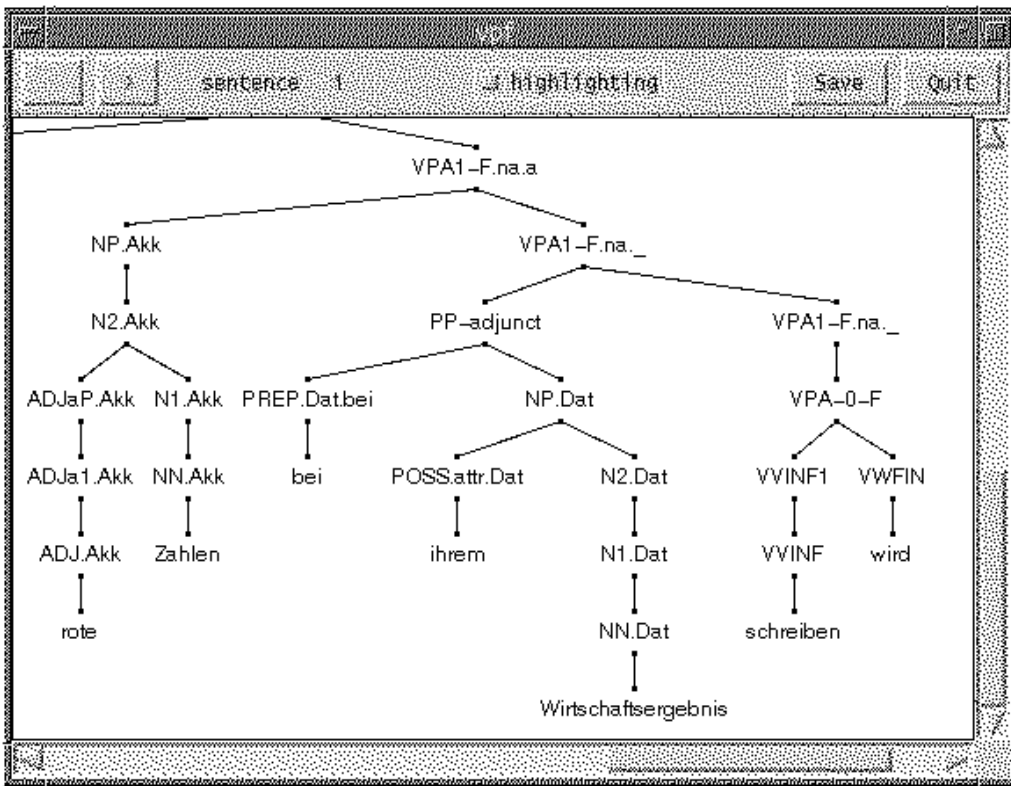


Figure 1: Detail of Viterbi parse: *Es wurde erwartet, dass die Flughafen Frankfurt AG (FAG) in diesem Jahr zum ersten Mal rote Zahlen bei ihrem Wirtschaftsergebnis schreiben wird.*

schreiben ‘be in the red’, lit. ‘red numbers write’. The accusative object is not adjacent to the verb but separated by the adjunct *bei ihrem Wirtschaftsergebnis* ‘at its economic result’.

- (1) Es wurde erwartet, dass die Flughafen Frankfurt AG (FAG) in diesem Jahr zum ersten Mal rote Zahlen bei ihrem Wirtschaftsergebnis schreiben wird .
 red figures at its economic result write will
 ‘It is expected that the Flughafen Frankfurt AG (FAG) will be in the red in its economic result this year for the first time.’

Figure 1 illustrates a Viterbi parse. A search routine collects the heads of all accusative objects together with the heads of their selecting verbs. In addition, it stores information about prenominal adjectival modifiers of the noun: either, it collects the head of the adjective or a mark which indicates that the noun is not modified, in which case the adjectival head feature is assigned the value ‘NoADJ’. We extracted only prenominal modifiers and ignored all postnominal modification.

For the extraction experiment, we used a corpus of 4,982,800 parsed newspaper sentences ranging from 5 to 30 words. We extracted 1,805,840 ANV-triples. 1,233,547 triples (about 68%) featured the feature ‘NoADJ’ instead of an adjectival modifier. After filtering potential parsing errors, i.e. triples with adjectives or verbs that were assigned the default lemma ‘unkown’, we ended up with 440,243 genuine triples, configurations in which an accusative object was modified by an adjectival modifier. About 70% of the modified occurrences are hapax legomena, i.e. triples that occurred only once in our corpus.

4 Calculation of Significance

The resulting quadruples, $\langle A, N, V, \text{frequency} \rangle$, are then evaluated by the log-likelihood ratio test (LL, Dunning 1993), a homogeneity test that compares the observed frequency of a pair of items with a frequency that is estimated under the assumption that the two items occurred independently of each other in the corpus, and that cooccurrence is a matter of chance. A high log-likelihood score means that the assumption of independence can be rejected with a high confidence and that it is probable that the pair is a significant combination. In particular, we compared the log-likelihood values of the three involved pairs: $\langle A, N \rangle$, $\langle N, V \rangle$, $\langle A, V \rangle$ ³, and furthermore the triple $\langle A, N, V \rangle$, which we simplified to the nested binary tuple $\langle \langle A, N \rangle, V \rangle$.

For the calculation of the log-likelihood ratio, we defined the probability space to consist of all observed triples $\langle A, N, V \rangle$ whereby we only considered prenominal attributive adjectives and accusative objects. This means that we ignore occurrences of the analyzed binary relations in other grammatical relations, e.g. whether a pair (A, N) occurred in subject function as well. We determined the log-likelihood scores of a tuple, e.g. $\langle A, N \rangle$, in dependence of the given triple $\langle A, N, V \rangle$ ignoring thereby all occurrences of V . This means we excluded all triples that included V from the probability space.

Sorting according to the log-likelihood ratio gives preference to significant combinations and suppresses random combinations of general highly frequent words. We implemented the ‘entropy version’ of log-likelihood, that makes reference to the partitions P_{ij} , rows R_i , and columns C_i of a contingency table and compares observed frequencies O_{ij} (read ‘observed frequencies O in partition P_{ij} ’) with expected frequencies E_{ij} (‘expected frequencies E in partition P_{ij} ’, cf. Evert 2002). For illustration, we give the contingency table for the calculation of the log-likelihood score of pair $\langle A, N \rangle$, given triple $\langle A, N, V \rangle$, in Table 1, taking the observed prenominal adjectives (plus the feature ‘NoAdj’) as one parameter (‘Adj’) and the accusative object nominal (‘Noun’) as the other. The equation in (2) shows how the log-likelihood score is determined from the information contained in the contingency table.

$$(2) \quad \text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

(Adj, Verb, Noun)	Noun	OtherNouns	
Adj	$O_{11} = \text{Adj, Noun, Other Verbs} $ $E_{11} = \frac{C_1 \cdot R_1}{N}$	$O_{12} = \text{Adj, OtherNouns, Other Verbs} $	$R_1 = O_{11} + O_{12}$
OtherAdjs	$O_{21} = \text{OtherAdjs, Noun, Other Verbs} $	$O_{22} = \text{OtherAdjs, OtherNouns, Other Verbs} $	$R_2 = O_{21} + O_{22}$
	$C_1 = O_{11} + O_{21}$	$C_2 = O_{12} + O_{22}$	$N = R_1 + R_2 = C_1 + C_2$

Table 1: Contingency table for LL(AN), given (Adj, Noun, Verb)

For each triple that was observed in the corpus, we collected four different log-likelihood scores. ‘LL’ abbreviates ‘log-likelihood score’. ‘A’, ‘N’, and ‘V’ are short forms of the involved constituents.

- (3) Given a triple (adjective, noun, verb) we calculate
- i. LL(ANV) whereby the normalizing factor is the set of all observed triples (including triples with the adjective feature ‘NoADJ’)

³We were pointed to the relevance of the pair $\langle A, V \rangle$ by Franz Josef Hausmann, p.c.

- ii. LL(AN) whereby the normalizing factor is the set of all observed triples to the exclusion of those triples that include the given verb
- iii. LL(NV) whereby the normalizing factor is the set of all observed triples to the exclusion of those triples that include the given adjective
- iv. LL(AV) whereby the normalizing factor is the set of all observed triples to the exclusion of those triples that include the noun at stake

The top 20 ANV-triples, sorted by LL(ANV), are displayed in table 4.

5 Preclassifying candidate sets

Sorting the resulting lists according to the different log-likelihood scores does only a partial job to discriminate the different classes. This is due to the fact that the log-likelihood scores of the different referent sets cannot be compared directly. Furthermore, due to the ‘binary treatment’ of the ternary word groups, the log-likelihood scores of the triples do not differentiate properly whether the involved pair (e.g. adjective, noun) is a significant pair as such or whether the two words are independent ‘outside’ the respective triple constellation. Ideally, we would expect the log-likelihood values of ANV, AN, and NV and their proportions to be correlated with the five classes we postulated in section 2, above. Table 2 summarizes these expected proportions⁴.

triple <A,N,V>	LL(ANV)	LL(AN)	LL(NV)
type i: ANV collocation	high	low	low
type ii: AN collocation	low	high	low
type iii: NV collocation	low	low	high
type iv: combination ii+iii	high	high	high
type v: trivial ANV	low	low	low

Table 2: Expected proportions of log-likelihood ratios (LL)

To approximate the intended classification, we employed an additional preprocessing means and trained a standard decision tree (C4.5, cf. Quinlan 1986) on a set of manually classified triples. Different versions of the decision tree helped to identify specific classes. We aimed at separating out, at least to a certain extent, idiomatic ANV-triples (type i), trivial combinations (type v) and the more strictly collocational cases (types ii, iii, and iv).

We obtained the best results by defining the decision tree attributes as relations between the different kinds log-likelihood values in combination with thresholds on the loglikelihood scores; furthermore we allowed the system to decide on additional thresholds on the frequency data. We implemented different versions of the decision tree based on (subsets of) the set of attributes listed in table 3.

The decision trees were trained on 89 manually classified examples and tested on 25 test examples, whereby the overall set of 114 examples was almost equally distributed over the five classes (25 from class i, 24 from class ii, 19 from class iii, 25 from class iv, and finally 21 from class v).

We did not find a decision tree which was able to discriminate all classes. Therefore, we decided to apply different runs of different decision trees to presort the data. Figure 2 gives the decision tree that performed best on class i items. All five class i examples were correctly classified as class i. There was only one false positive, a class iii item falsely identified as class i. All in all, it produced 13 errors (52.0%) on the 25 test examples: the other classes were not discriminated as well as class i. The tree is given in standard C4.5 notation. The number to the right of a colon denotes the classification, the first number in round brackets to the right names the number of times the path was followed in the

⁴We disregard the log-likelihood score of pairs of adjective+verb, here, since this combination is not collocational in the syntactic contexts we analyze.

attribute	condition	value	else
A	$ll(anv) > ll(an)$	y	n
B	$ll(anv) > ll(nv)$	y	n
C	$ll(anv) > ll(av)$	y	n
D	$ll(an) < 100$ and $ll(nv) < 100$	y	n
E	$ll(an) < 100$	y	n
F	$ll(nv) < 100$	y	n
G	$ll(an) - ll(nv) < ll(an)/2$	y	n
H	$ll(nv) - ll(an) < ll(nv)/2$	y	n
I	$ll(anv) > 50$ and $ll(av) < 10$	y	n
J	$ll(an) > ll(nv)$	y	n
attributes $ll(anv)$, $ll(an)$, $ll(nv)$, $ll(av)$, $f(anv)$, $f(a)$, $f(n)$, $f(v)$, $f(an)$, and $f(nv)$		with continuous values	

Table 3: Attributes of decision tree learning

training. The optional number to the right of this determines the number of errors made at this point in the decision tree during training. The decision tree classifies 58 triples as elements of class i.

6 Results

In section 2, we described five different types of ANV-triples. In our experiment on automatically extracting those triples from a newspaper corpus, we used a stochastic grammar, sorting by means of the log-likelihood ratio and clustering by means of a decision tree trained on a set of manually classified data.

In table 4, the 20 best ANV-triples are given, sorted by the log-likelihood scores $LL(ANV)$ of the triples. The log-likelihood scores are less clearly interpretable than idealized in table 2, above: this is evident from the log-likelihood figures provided alongside the manually classified candidates given in table 5 (same table layout), which contains a certain number of hapax legomena and low frequency items.

A	N	V	$LL(ANV)$	$f(ANV)$	$LL(AN)$	$LL(NV)$	$LL(AN)$
groß	Rolle	spielen	4898.20	486.00	38.80	32585.80	6.05
wichtig	Rolle	spielen	4152.43	431.00	505.91	26387.16	0.18
schwer	Verletzung	erleiden	3358.58	314.00	1112.44	2375.46	1591.45
technisch	Entwicklung	aufzeigen	2883.93	187.00	83.86	50.98	3.90
leicht	Verletzung	erleiden	2747.62	241.00	528.69	2897.80	677.77
rot	Zahl	schreiben	2070.60	192.00	555.51	1172.95	2.56
schwarz	Zahl	schreiben	2067.46	185.00	305.23	778.12	3.72
entscheidend	Rolle	spielen	1827.06	190.00	211.57	25779.82	0.15
offen	Tür	einrennen	1726.96	94.00	95.18	200.67	6.58
entsprechend	Beschluß	fassen	1709.36	134.00	427.11	2986.22	22.29
grün	Licht	geben	1622.59	375.00	1398.28	0.15	43.87
klar	Absage	erteilen	1522.36	120.00	123.76	2854.22	49.06
heftig	Kritik	üben	1453.40	126.00	852.48	4334.97	7.87
ordnend	Rolle	spielen	1303.58	130.00	43.88	9843.82	0.00
groß	Wert	legen	1292.30	113.00	51.14	3924.96	4.84
eigen	Weg	gehen	1130.17	112.00	200.95	4516.22	1.32
schwer	Vorwurf	erheben	1121.08	98.00	320.92	982.47	34.92
neu	Weg	gehen	1115.76	129.00	845.40	4495.05	1.02
positiv	Bilanz	ziehen	1069.24	113.00	722.88	2187.01	246.32

Table 4: Results sorted by log-likelihood ratio $LL(ANV)$

```

ll(anv) <= 265.1 :
| f(a) > 5854 : 3 (7.0/1.0)
| f(a) <= 5854 :
| | B = y:
| | | c = n: 4 (4.0/1.0)
| | | C = y:
| | | | G = y: 5 (2.0)
| | | | G = n:
| | | | | ll(nv) > 6.34 : 2 (2.0)
| | | | | ll(nv) <= 6.34 :
| | | | | | f(a) <= 205 : 2 (6.0/1.0)
| | | | | | f(a) > 205 : 5 (6.0)
| | B = n:
| | | H = y:
| | | | f(a) <= 923 : 4 (2.0)
| | | | f(a) > 923 : 2 (3.0)
| | | H = n:
| | | | f(a) <= 417 : 4 (6.0/2.0)
| | | | f(a) > 417 :
| | | | | f(nv) <= 43 : 3 (2.0)
| | | | | f(nv) > 43 : 5 (3.0/1.0)
ll(anv) > 265.1 :
| f(nv) <= 390 :
| | A = n: 2 (2.0)
| | A = y:
| | | ll(av) <= 4.81 :
| | | | f(n) <= 473 : 1 (2.0)
| | | | f(n) > 473 : 2 (5.0/1.0)
| | | ll(av) > 4.81 :
| | | | B = y:
| | | | | ll(nv) <= 222.98 : 1 (13.0)
| | | | | ll(nv) > 222.98 : 4 (3.0/1.0)
| | | | B = n:
| | | | | ll(an) <= 138.25 : 1 (3.0)
| | | | | ll(an) > 138.25 : 3 (2.0)
f(nv) > 390 :
| | f(v) <= 4925 :
| | | f(nv) > 770 : 4 (4.0)
| | | f(nv) <= 770 :
| | | | ll(av) <= 48.18 : 3 (4.0)
| | | | ll(av) > 48.18 : 4 (2.0)
| | f(v) > 4925 :
| | | ll(anv) <= 442.72 : 5 (3.0)
| | | ll(anv) > 442.72 : 4 (3.0)}

```

Figure 2: Decision Tree that performed best for class i

Table 4 contains several combinations with the noun *Verletzung* ‘injury’, which have high frequency figures: this may be an artefact of our newspaper corpus. The result data also include examples which are not fully captured by the five descriptive types of Section 2. These are cases in which an adjective is required but not restricted to a specific lexical item like *<ordinal number> Lebensjahr vollenden* ‘complete the nth year of life’.

Potentially non-collocative triples (type v) are likely to contain general, non-collocative adjectives like the deictic *entsprechend* ‘corresponding’ or the listing item *weiter* ‘further’. We have heuristically extracted some of these, setting thresholds such that the $LL(ANV) < 20$, $LL(AN) < 30$, $LL(AV)$, and $F(A) > 1000$. This gives a list of 70 adjectives many of which satisfy this criterion, cf. (4) for a sample. Such adjectives and the pertaining ANV-triples would be removed from the material to be given to a lexicographer for manual subclassification.

- (4) *ander, besonder, bestimmt, deutlich, eigene, einzig, entscheidend, entsprechend, erheblich, folgend, ganz, gesamt, gleich, gut, sogenannt, ...*

Table 6 contains the top ten of the ANV-triples classified as belonging to type i (left column) and the top ten from type v (right column). In the left column, we have marked with an asterisk (*) those

type	A	N	V	LL(ANV)	Freq(ANV)	LL(AN)	LL(NV)	LL(AV)
type i	offen	Tür	einrennen	1726.96	94	95.18	200.67	6.58
	golden	Nase	verdienen	679.78	55	68.39	0.91	59.76
	kalt	Schulter	zeigen	415.87	41	0.00	33.31	11.29
	letzt	Wort	haben	404.46	124	404.12	5.98	10.07
type ii	rot	Zahl	schreiben	2070.60	192	555.51	1172.95	2.56
	absolut	Mehrheit	verlieren	315.65	62	4667.07	205.12	0.00
	gut	Ruf	genießen	258.78	29	1045.71	384.04	1.20
	offen	Drogenszene	entgegenwirken	13.01	1	260.03	0.00	0.00
	einstweilig	Verfügung	erwirken	606.67	44	911.66	3.38	73.55
	diplomatisch	Beziehung	aufnehmen	551.21	66	1142.49	115.06	4.81
	archimedisch	Punkt	lokalisieren	19.41	1	21.11	0.00	0.00
	Rot	Liste	ansehen	12.04	1	27.15	0.00	0.00
type iii	entsprechend	Beschluß	fassen	1709.36	134	427.11	2986.22	22.29
	neu	Arbeitsplatz	schaffen	337.58	48	261.70	1216.29	762.40
	bestehend	Verlustvortrag	tilgen	19.62	1	0.00	34.41	0.00
	gemeinsam	Vorstellung	entwickeln	8.75	1	2.87	106.03	169.02
type iv	scharf	Kritik	üben	2432.86	199	704.72	3208.22	7.30
	klar	Absage	erteilen	1522.36	120	123.76	2854.22	49.06
	dringend	Appell	richten	200.48	14	7.43	758.98	41.70
	einstimmig	Urteil	fällen	12.64	1	1.91	640.37	9.84
type v	konkret	Zahl	nennen	423.27	51	166.85	1509.10	143.72
	neu	Gast	begrüßen	19.41	2	0.79	1235.58	307.36
	alt	Eiche	entwurzeln	18.28	1	40.11	0.00	0.00
	gesamt	Film	durchziehen	12.53	1	21.72	32.27	57.95

Table 5: Manually classified examples

items which we would manually classify as belonging to type i; the other items belong to type iv.

offen Tür einrennen *	konkret Zahl nennen
deutlich Sprache sprechen *	ander Problem haben
frei Lauf lassen *	deutlich Zeichen setzen
klein Brötchen backen *	weit Auskunft geben
golden Nase verdienen *	genau Angabe machen
reißend Absatz finden	fatal Folge haben
groß Aufsehen erregen	groß Sorge machen
schwer Geschütz auffahren *	groß Schwierigkeit haben
groß Anklang finden	weit Information geben
gut Haar lassen *	aufschiebend Wirkung haben

Table 6: ANV-triples classified by the decision tree: top ten from type i (left) and from type v (right)

7 Discussion and Outlook

Our objective is to provide raw data on significant ANV-triples for lexicographers; in addition, these data are to be sorted in a way that should allow the lexicographers to manually evaluate the data with little effort.

We argue that preprocessing based on both, linguistic knowledge and statistical information, is superior to shallow methods or simple part-of-speech pattern matching. The probabilistic grammar allows us to also identify non-adjacent configurations and to thus produce raw candidate data which are homogeneously of the same syntactic type.

Statistical sorting by means of the log-likelihood ratio test helps to identify significant triples and to even out the impact of general high frequency items. To improve the results of this test and to make the figures more easily comparable, the current pair-based calculation of association measures would need to be extended to word triples. Nevertheless, the proportions between the log-likelihood ratios of the ANV-triples and of the (possibly related) NA and NV collocations seem to constitute a starting point for further subclassifying the data into idiomatic vs. collocational vs. trivial.

8 References

Bouma, Gosse, and Begoña Villada. 2002. Corpus-based acquisition of collocational prepositional phrases. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19:1:61–74.

Moira Runcie et al. 2002. *OCDSE – Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Evert, Stefan. 2002. Mathematical Properties of AMs. Handout, Workshop Computational Approaches to Collocations, Vienna.

Evert, Stefan, and Brigitte Krenn. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Firth, John Rupert. 1957. *Studies in linguistic analysis*, chapter A synopsis of linguistic theory 1930–55, pp. 1–32. Oxford.

Hausmann, Franz Josef 2003. Was sind eigentlich Kollokationen? Talk at IDS Jahrestagung, to appear.

Heid, Ulrich. 1994. On Ways Words Work Together – Topics in Lexical Combinatorics. In Willy Martin et al. (eds.), *Proceedings of the VIth Euralex International Congress*, pp. 226 – 257, Amsterdam.

Krenn, Brigitte. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*. PhD thesis, DFKI and Universität des Saarlandes, Saarbrücken.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, 1st edition. Cambridge (MA): MIT Press.

Quinlan, John Ross. 1986. Induction of Decision Trees. *Machine Learning* 1:81–106.

Schmid, Helmut. 2000. Lopar: Design and Implementation. Arbeitspapiere des Sonderforschungsbereichs 340 *Linguistic Theory and the Foundations of Computational Linguistics* 149, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schulte im Walde, Sabine. 2003. A collocation database for german verbs and nouns. Budapest. COMPLEX 2003.

Schulte im Walde, Sabine, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. 2001. Statistical Grammar Models and Lexicon Acquisition. In Christian Rohrer, Antje Rossdeutscher, and Hans Kamp (eds.), *Linguistic Form and its Computation*, pp. 387–440. Stanford, CA: CSLI Publications.