Between complex predicates and regular phrases: some collocational combinations in German

Philippa Cook, Freie Universität Berlin

1   Introduction

I take a new approach to certain multi-word strings in German such as (i) *heftig in die Kritik geraten* 'to be heavily criticised', (ii) *immer der Erste sein* 'to always be the first', (iii) *richtig Geld verdienen* 'to make real money' or (iv) *richtig Gas geben* 'increase effort'. The string in (i) involves a (semi-compositional) support verb construction (or FVG cf. Krenn and Erbach (1994), Steinitz (1989)) *in die Kritik geraten* (literally: into the criticism fall) further modified by a modifier *heftig* 'harshly' (an adverbally used adjective). The multi-word string in (ii) involves the copula *sein* 'be' with a nominal predicative phrase *der Erste* 'the first one'. There is further modification by *immer* 'always'. In (iii) a verb *verdienen* selects a bare noun, and there is modification by an adverbally used adjective *richtig*. This string is fairly compositional in contrast to strings such as (iv), involving the bare noun *Gas* combined with the verb *geben*, which have an idiomatic meaning. Common to each of the multi-word strings is that we have a modifier, a noun or a PP, and a verb. I offer a novel analysis of such strings, postulating units I call (*modifier*)-*collocational chunks*.

In the analysis, I take these lexical strings to be exemplific of several larger classes of data patterning similarly, although with small differences across subclasses. For reasons of space I cannot document the full array of data here but in the paper I will introduce a larger body of data (cf. also "Outlook" below).

I will show that there is collocational relationship (cf. Firth (1957), Sinclair (1991, 1996), Evert (2008)) spanning all three subcomponents of the string and I will go a step further and argue that such expressions may combine in syntax in German via a special schema for building collocational chunks (rather than building traditionally known syntactic constituents). The schema I propose is inspired by Function Composition as used in Categorial Grammar. Each of the three elements in the string is individually a syntactic atom of a multi-word string. The combination as a whole should be considered one complex lexeme, the building of which in syntax is licensed lexically.

First I show the collocational nature of the subcomponents of the tuple. I show that (i) the modifier collocates with the noun, (ii) the noun collocates with the verb and (iii) the modifier collocates with the verb. They are thus "tuples" - three-word collocational word groups - similar (though different) to those discussed by Zinsmeister and Heid (2003).

2   The collocational nature of the data

2.1   Wortprofil measures

The Wortprofil 3.0 tool of the DWDS (cf. Didakowski and Geyken (2013)) corpus was used to show that the expressions are collocate-tuples.

Association Measures for *heftig in die Kritik geraten*

lexical string	association using LogDice	frequency
heftig + Kritik	11.12	9882
geraten + in Kritik	8.96	3843
Kritik + geraten-in	9.27	2453

Association Measures for *richtig Geld verdienen*

lexical string	association using LogD	frequency
richtig + Geld	6.97	68
Geld + verdienen	11.51	22226
richtig + verdienen	6.09	332

A comprehensive study of the collocational behaviour of these tuples would warrant a separate paper, for now my aim is simply to provide a glimpse at the empirical basis underlying the syntactic analysis. Further measures will be provided in the full paper.

3 A Syntactic Analysis for Collocational Strings

3.1 Multiple Fronting data as an indicator of combinatoric possibility

Semantically the modifier seems to modify the whole PP/N + V string in the data in question. Syntactically, though, we find constructions where the modifier (surprisingly) combines with the PP or the N. These are so-called multiple fronting constructions where the clause-initial position before the finite verb (the "front field") – which can only house a constituent – contains the modifier + N/PP string. We have a non-isomorphism of syntax and semantics, as well as a curious constituent structure (see Müller (2003, 2005)).

(1) a. **Richtig Geld** wird nur im Briefgeschäft verdient[1]
       really   money is   only in lettercommerce made

       'You can only make big money with letters'

   b. **Heftig in die Kritik** geriet der Kostenrechnungsbericht des Jugendamtes für 2002[2]
      heavily into the criticism fell   the finance report   the youth service for 2002

       'The youth service's 2002 financial report got heavily criticised'

It is important in analysing multiple fronting constructions of this type that we take into account not only the tight bond between the elements in the front field but also the tight bond between the elements in the front field and the syntactically separate verb. Also note that the string in initial position above (e.g. *heftig in Kritik*) can be scrambled within the middlefield as a complex unit. The lack of inflection makes clear we are not dealing with an adjectivally modified noun here.

(2) weil   **heftig in Kritik** der Kostenrechnungsbericht des Jugendamtes geraten ist
    because heavy in criticism the finance report   the youth services fallen   is

    'because the youth service's 2002 financial report got heavily criticised '
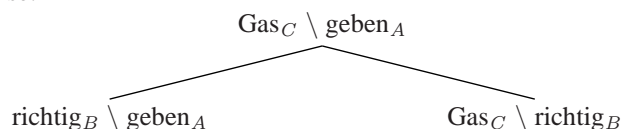
Evidence against assuming one phrasal lexical entry for the string concerns the syntactic mobility of subparts of the strings and passive (cf. (1a) above). Atoms of the complex lexeme display a degree of syntactic mobility higher than that of members of genuine complex predicates but lower than that of regularly composed syntactic phrases.

3.2 Transferring Function Composition to the collocational domain

A solution to the syntax-semantics mismatch is the use of Function Composition (FC) instead of Functional Application to combine elements in syntax (cf. Jacobson (1990)). FC combines two functors to yield a new functor as sketched here:

(3) Forward Function Composition A/B $*$ B/C = A/C

(4) Backward Function Composition B\A $*$ C\B = C\A

Informally speaking, taking the string *richtig Gas geben* [lit. really give gas] 'increase effort' as an example: *richtig* is looking for *geben* as a collocate (and as its modifiee) and *Gas* is looking for *richtig* as a collocate too. The 'need' for *geben* is postponed to the next level. I sketch here how the FC of *richtig* and *Gas* can be interpreted in a backward FC sense:

$$\text{Gas}_C \setminus \text{geben}_A$$

$$\text{richtig}_B \setminus \text{geben}_A \qquad\qquad \text{Gas}_C \setminus \text{richtig}_B$$

I define the following lexical entries for *richtig Gas geben* 'increase effort'

---

[2]taz, 28-29.10.2000, p. 5

[2]COSMAS, RHZ03/SEP.09166 Rhein-Zeitung, 12.09.2003

Lexical entry for *richtig*

$$
\begin{bmatrix}
word \\
\text{PHON} & \langle richtig \rangle \\
\text{SS|LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{MOD} & V\begin{bmatrix} \text{LID} & \boxed{4} \end{bmatrix} \\ \text{LID} & \textit{richtig-intensifier} \end{bmatrix} \\ \text{SUBCAT} & \langle \rangle \end{bmatrix} \end{bmatrix} \\
\text{CONT} & \begin{bmatrix} \textit{intensify} & \boxed{4} \end{bmatrix} \\
\text{COLL|LID} & \textit{geben-idiomatic} \; \boxed{4}
\end{bmatrix}
$$

Lexical entry for *Gas*-$_{idiomatic}$

$$
\begin{bmatrix}
word \\
\text{PHON} & \langle Gas \rangle \\
\text{SS|LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{LID} & \textit{Gas-idiomatic} \end{bmatrix} \\ \text{SUBCAT} & \langle \rangle \\ \text{SPR} & \langle \rangle \end{bmatrix} \end{bmatrix} \\
\text{CONT} & \begin{bmatrix} \text{INDEX} & \textit{non-referential} \end{bmatrix} \\
\text{COLL|LID} & \textit{richtig-intensifier}
\end{bmatrix}
$$

Lexical entry for the light verb *geben*-$_{idiomatic}$

$$
\begin{bmatrix}
word \\
\text{PHON} & \langle geben \rangle \\
\text{SS|LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{LID} & \textit{geben-idiomatic} \end{bmatrix} \\ \text{SUBCAT} & \langle \text{NP-NOM} \; \boxed{5} \rangle \end{bmatrix} \end{bmatrix} \\
\text{CONT} & \begin{bmatrix} \text{RELS} & \left\langle \begin{bmatrix} \textit{increase effort} \\ \text{AGENT} & \boxed{5} \end{bmatrix} \right\rangle \end{bmatrix} \\
\text{COLL|LID} & \textit{Gas-idiomatic}
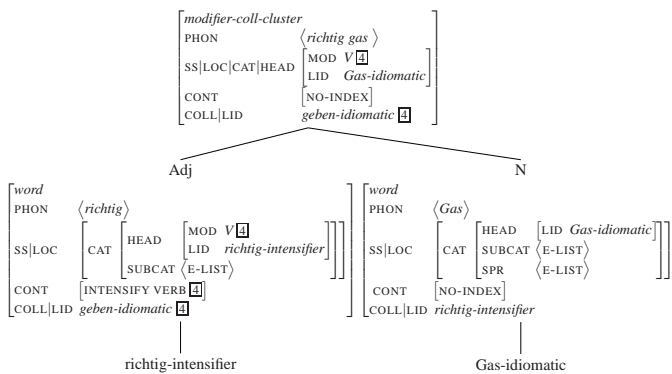\end{bmatrix}
$$

The LID (lexical identifier) feature appropriate for the sort *head* (cf. Richter and Sailer (1999); Soehn (2004), also Sag (2012); Spencer (2005)) identifies specific instantiations of words. The COLL feature (*ibid*) (here appropriate for type *word*) encodes in the lexical entry that it collocates with a particular lexeme. The lexical entry of the modifier specifies that the modified phrase is the verb. Here we see the encoding of the syntax-semantics mismatch as *richtig* modifies the verb semantically but combines (in this collocational environment) syntactically with the noun first.

The schema for licensing the *modifier-collocational-cluster richtig Geld* is given here

modifier-coll-cluster $\rightarrow$

$$
\begin{bmatrix}
\text{SS} & \begin{bmatrix} \text{LOC|CAT|HEAD} & \begin{bmatrix} \text{MOD} & \boxed{4} \\ \text{LID} & \boxed{1} \end{bmatrix} \\ \text{COLL|LID} & \boxed{4} \\ \text{CONT} & \boxed{6} \end{bmatrix} \\
\text{NON-HD-DTR} & \begin{bmatrix} \text{SS|LOC|CAT} & \begin{bmatrix} \text{HEAD|MOD} & \boxed{4} \\ \text{LID} & \boxed{2} \end{bmatrix} \\ \text{COLL|LID} & \boxed{4} \end{bmatrix} \\
\text{HD-DTR} & \begin{bmatrix} \text{SS|LOC|CAT} & \begin{bmatrix} \text{HEAD|LID} & \boxed{1} \\ \text{SPR} & \langle \rangle \end{bmatrix} \\ \text{COLL|LID} & \boxed{2} \\ \text{CONT} & \boxed{6} \end{bmatrix}
\end{bmatrix}
$$

At the cluster (MTHR) level, just the verb is required via MOD and is collocationally required too. The subtree below illustrates the cluster *richtig Gas* as licensed by the above schema:

Finally, I will turn to the combination of the *mod-coll-chunk* with the verb. This can be licensed by the *head-adjunct-schema*. Semantic modification of the verb by *richtig* occurs there.

```
⎡ modifier-coll-cluster                              ⎤
⎢ PHON              ⟨ richtig gas ⟩                   ⎥
⎢                  ⎡ MOD   V ▣ ⎤                      ⎥
⎢ SS|LOC|CAT|HEAD  ⎢ LID   Gas-idiomatic ⎥           ⎥
⎢                  ⎣           ⎦                      ⎥
⎢ CONT              [ NO-INDEX ]                      ⎥
⎣ COLL|LID          geben-idiomatic ▣                ⎦
                    /                        \
                  Adj                          N

⎡ word                                        ⎤   ⎡ word                                              ⎤
⎢ PHON   ⟨ richtig ⟩                          ⎥   ⎢ PHON      ⟨ Gas ⟩                                 ⎥
⎢        ⎡      ⎡ MOD  V ▣           ⎤⎤        ⎥   ⎢           ⎡        ⎡ HEAD  [ LID  Gas-idiomatic ]⎤⎤ ⎥
⎢ SS|LOC ⎢ CAT  ⎢ HEAD ⎢ LID  richtig-intensifier ⎥⎥⎥   ⎢ SS|LOC   ⎢ CAT   ⎢ SUBCAT ⟨ E-LIST ⟩ ⎥⎥ ⎥
⎢        ⎢      ⎣ SUBCAT ⟨ E-LIST ⟩ ⎦⎦        ⎥   ⎢           ⎣        ⎣ SPR    ⟨ E-LIST ⟩ ⎦⎦ ⎥
⎢ CONT   [ INTENSIFY VERB ▣ ]                 ⎥   ⎢ CONT      [ NO-INDEX ]                            ⎥
⎣ COLL|LID  geben-idiomatic ▣                 ⎦   ⎣ COLL|LID  richtig-intensifier                    ⎦
              |                                              |
         richtig-intensifier                           Gas-idiomatic
```

4 Outlook: Extension to more and less abstract tuples

The lexical string I examined here has three lexically fixed atoms. As mentioned above, there is reason to extend the current analysis to numerous other classes of tuples in German that display similar syntactic behaviour. Some of these tuples have slots that can be instantiated by a certain class of lexeme, e.g.(i) am ADJ-sten + directional PP + motion verb, e.g. *am billigsten in die Türkei kommen* 'to get to Turkey the cheapest way'or (ii) 'free' dative pronoun + PP + state verb, e.g. *ihm zur Seite stehen* 'to stand (to his benefit) by his side'. The data span a continuum invloving intermediate tuples where one or two of the exponents are lexically fixed. At the most extreme end of the spectrum lie the fixed idioms with non-compositional meaning such as *den Stein ins Rollen bringen* 'get the ball rolling'. Thus, my proposed analysis will need to be modified to account for this variation in abstractedness of the strings (lexically fixed vs. slots for lexeme classes). Further, I will show how a variant of the schema can be used to capture instances of Integration of objects and (certain) subjects in the sense of e.g. Jacobs (1993). I also propose to handle extended copula constructions such as *von Experten geprüft sein* [lit. from experts verified be] 'be expert-checked' (see Maienborn (2011)) using a cluster schema for combination of *von Experten* with *geprüft*. Thus, the body of data covered by cluster analyses will be quite large.

5 Summary and Conclusions

I argue for complex multi-word expressions consisting of collocational tuples to be lexically licensed in German. I propose a schema licensing the syntactic building of non-standard constituents ("chunks") comprising a modifier and a noun/PP. This schema can license the non-standard constituent found in multiple fronting constructions. Since these units are only licensed in very lexically specific environments, I expect their occurrence to be highly restricted (and in fact multiple fronting is rare). Jacobson (1990) claimed that FC should not apply freely across the grammar but only in special cases and I also believe that this type of non-canonical syntactic combination only occurs when lexically licensed by the collocational nature of the material involved. Although I license the building of this complex string in the syntax, its actual licensing is lexical in the sense that it is sanctioned by the lexical entries. I show how I consider units such as *richtig Geld*, *Geld verdienen*, and *richtig Geld verdienen* to be lexically licensed collocations that may form (non-canonical) syntactic phrases. These are thus very close to what have been called "exemplars" or "chunks" in the usage-based literature (cf. e.g. Beckner and Bybee (2009)). Motivating and formally encoding such non-standard constituents and meshing the formalism with frequency and usage data is an interesting development. Finally, the research reported on here has ramifications for the argument/adjunct distinction and introduces the notion of collocationally-selected modifiers as intermediate between arguments and pure modifiers in the traditional sense (cf. Dowty (2003)).

References

Beckner, Clay and Bybee, Joan. 2009. A usage-based account of constituency and reanalysis. *Language Learning* 59, 27–46.

Didakowski, Jörg and Geyken, Alexander. 2013. From DWDS corpora to a German Word Profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network "Internet Lexicography"*, OPAL - Online publizierte Arbeiten zur Linguistik X/2012, pages 43 – 52, Mannheim: Institut für deutsche Sprache.

Dowty, Davis. 2003. The Dual Analysis of Adjuncts and Complements in Categorial Grammar. In Ewald Lang, Claudia Maienborn and Cathrine Fabricius-Hansen (eds.), *Modifying Adjuncts*, pages 1 – 22, de Gruyter.

Evert, Stefan. 2008. Corpora and Collocations. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics. An International Handbook. Article 58*, pages 1212 – 1248, Berlin: Mouton de Gruyter.

Firth, J. R. 1957. A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis*, pages 1 – 32, Oxford: The Philological Society.

Jacobs, Joachim. 1993. Integration. In Marga Reis (ed.), *Wortstellung und Informationsstruktur*, page 63–116, Tübingen: Niemeyer.

Jacobson, Pauline. 1990. Raising as Function Composition. *Linguistics and Philosophy* 13, 423–475.

Krenn, Brigitte and Erbach, Gregor. 1994. Idioms and Support Verb Constructions. In John Nerbonne, Klaus Netter and Carl J. Pollard (eds.), *German in Head-Driven Phrase Structure Grammar*, CSLI Lecture Notes, No. 46, pages 365–396, Stanford: CSLI Publications.

Maienborn, Claudia. 2011. Strukturausbau am Rande der Wörter: Adverbialer Modifikatoren beim Zustandspassiv. In Stefan Engelberg et. al (ed.), *Sprachliches Wissen zwischen Lexikon und Grammatik*, pages 317 – 343, Berlin: de Gruyter.

Müller, Stefan. 2003. Mehrfache Vorfeldbesetzung. *Deutsche Sprache* 31(1), 29–62.

Müller, Stefan. 2005. Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. *Linguistische Berichte* 203, 297–330.

Richter, Frank and Sailer, Manfred. 1999. LF Conditions on Expressions of Ty2: An HPSG Analysis of Negative Concord in Polish. In Robert D. Borsley and Adam Przepiórkowski (eds.), *Slavic in Head-Driven Phrase Structure Gram- mar*, pages 247 – 282, Stanford: CSLI.

Sag, Ivan. 2012. Sign-based Construction Grammar: An informal Synopsis. In Hans Boas and Ivan Sag (eds.), *Sign-based Construction Grammar*, pages 69 – 202, Stanford: CSLI.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford university Press.

Sinclair, John. 1996. Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday and R. H. Robins (eds.), *Im Memory of J. R. Firth*, pages 410 – 430, London: Longman.

Soehn, Jan-Philipp. 2004. License to Coll: How to Bind Bound Words and Readings to their Contexts. In Stefan Müller (ed.), *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pages 261 – 273, Stanford: CSLI.

Spencer, Andrew. 2005. Generalized Paradigm Function Morphology – A synopsis. In Alexandra Galani and Beck Sinar (eds.), *York Papers in Linguistics*, volume 2 of *Papers from the York-Essex Morphology Meeting, 2003*, pages 93 – 106, University of York, UK.

Steinitz, Renate. 1989. Überlegungen zum Prädikativ. In Wolfgang Motsch (ed.), *Wortstruktur und Satzstruktur*, pages 211 – 234, Berlin: Zentralinstitut für Sprachwissenschaft.

Zinsmeister, Heike and Heid, Ulrich. 2003. Significant triples: Adjective + noun + verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research*, COMPLEX 2003, pages 92 – 101, Budapest, Hungary.