Counting Markedness.
A corpus investigation on German free relative constructions

Ralf Vogel *Marco Zugck*[*]

University of Potsdam

This paper reports the results of a corpus investigation on case conflicts in German argument free relative constructions. We investigate how corpus frequencies reflect the relative markedness of free relative and correlative constructions, the relative markedness of different case conflict configurations, and the relative markedness of different conflict resolution strategies. Section 1 introduces the conception of markedness as used in Optimality Theory. Section 2 introduces the facts about German free relative clauses, and section 3 presents the results of the corpus study. By and large, markedness and frequency go hand in hand. However, configurations at the highest end of the markedness scale rarely show up in corpus data, and for the configuration at the lowest end we found an unexpected outcome: the more marked structure is preferred.

1 Markedness in OT

In Optimality Theory, grammaticality is derived from markedness in the sense that it is the relative ranking of markedness constraints that determines whether a structure is grammatical or not. Consider the following simple system of two

[*] The division of labour among the authors was as follows: Zugck carried out the low-level work on the corpus, data sample extraction, counting, systematising the numerical results, some calculations. The higher level linguistic analysis was done by Vogel.

Linguistics in Potsdam 21 (2003): 105–122
Susann Fischer, Ruben van de Vijver, Ralf Vogel (eds.)
Experimental Studies in Linguistics 1

markedness constraints **M1** and **M2**, one faithfulness constraint **F**, and two candidates *cand1* and *cand2*:

(1)

	M1	**M2**	**F**
cand1		∗	
cand2	∗		

The input either conforms to *cand1* or *cand2*. Constraint **F** favours the candidate referred to in the input. Assume further that the relative ranking of **M1** and **M2** is universally fixed, which is typical for two markedness constraints that express a markedness scale. Under these circumstances, *cand1* is grammatical (i.e., the winner of at least one OT competition) under any possible ranking, while the grammaticality of *cand2* depends on the relative ranking of **F**. The four tables in (2) show this:

(2) a. *A grammar with low-ranked faithfulness*

cand1	**M1**	**M2**	**F**
☞*cand1*		∗	
cand2	∗!		∗

	cand2	**M1**	**M2**	**F**
☞	*cand1*		∗	∗
	cand2	∗!		

b. *A grammar with high-ranked faithfulness*

cand1	**F**	**M1**	**M2**
☞*cand1*			∗
cand2	∗!	∗	

cand2	**F**	**M1**	**M2**
cand1	∗!		∗
☞*cand2*		∗	

The following observations concerning the relative markedness of *cand1* and *cand2* can be made:

- the set of languages where *cand2* is grammatical, is a subset of those where *cand1* is grammatical
- In order to be grammatical, *cand2* needs highly ranked faithfulness

These observations are indicative of the higher markedness of *cand2*. A third observation that can often be made is that for those languages where the more marked *cand2* is possible, the set of contexts in which it occurs is a subset of the contexts where *cand1* is possible.

What are the empirical predictions of such a model of markedness? In grammaticality judgement tasks, we expect that *cand2* is more likely to be judged as ungrammatical than *cand1*, at best as equal, but never better. For research on corpora, we expect higher frequencies of the less marked expressions. Section 2 introduces the case of German free relative clauses that realise an argument of the verb. The relation of free relative clauses and correlative clauses in German is an instructive example for the kind of markedness relation just discussed. Section 3 reports the results of a corpus investigation on this construction.

2 German Argument Free Relative Constructions

Vogel (2001, 2002) showed that argument free relative (FR) constructions in German display tendencies of markedness in various ways. The first observation is that FR constructions are marked as such. The FR pronoun has to serve two case assigners at the same time:

(3) Wer sich nicht wehrt, lebt verkehrt
Who-NOM SELF not defends lives wrongly

In this example, '*wer*' is the subject of the underlined FR clause, and the whole FR is the subject of the matrix clause. Both finite verbs assign nominative case to their subject, but there is only one element, the FR pronoun, that realises nominative case. FRs as such are marked syntactic constructions. There are languages that do not have FR constructions in the way exemplified in (3), for instance, Hindi (Dayal, 1996) and Korean (Vogel, 2000). In those languages, a FR is typically left dislocated and 'doubled' by a correlate pronoun. This 'correlative' construction (CORR) is also always possible in languages with FRs. The correlative counterpart of (3) is (4):

(4) Wer sich nicht wehrt, der lebt verkehrt
Who-NOM SELF not defends that-one-NOM lives wrongly

Vogel (2000) suggested a markedness constraint 'case uniqueness' (CU) that requires a one-to-one relation between case assigners and case assignees. FRs violate this constraint. Hence, they only survive, if faithfulness is ranked higher than this constraint:

(5) a. *Languages without FRs*

	FR	CU	F
	FR	∗!	
☞	CORR		∗

	CORR	CU	F
	FR	∗!	∗
☞	CORR		

b. *Languages with FRs*

	FR	F	CU
☞	FR		∗
	CORR	∗!	

	CORR	F	CU
	FR	∗!	∗
☞	CORR		

Languages with FRs further differ in the way they realise FRs, in particular, we find three different kinds of strategies that differ in which case is realised, the case assigned by the matrix verb (`m-case`) or by the relative clause internal verb (`r-case`), and how:

Strategy M: The FR pronoun realises `m-case`

Strategy R: The FR pronoun realises `r-case`

Strategy RES: The FR pronoun realises `m-case`, and is accompanied by a resumptive pronoun realising `r-case`

German FRs always use strategy R, Icelandic ones strategy M (Vogel, 2000), Gothic (Harbert, 1983) and Romanian (Grosu, 1994) shift between the two options depending on which case is more prominent on the language's case hierarchy. Modern Greek (Alexiadou and Varlokosta, 1995) uses strategy M, and strategy RES, if `m-case` is structural, and `r-case` oblique. See (Vogel, 2000,

2002) for a detailed discussion of the typology of case conflicts in argument FR constructions.

Given the fact that pronouns can realise only one case, this configuration becomes problematic, whenever the two cases differ. English (Bresnan and Grimshaw, 1978) and Dutch (Groos and van Riemsdijk, 1981) are reported to be 'matching' languages – they only allow for FRs if the two cases match.

German has also been reported to be a matching language (Groos and van Riemsdijk, 1981). But this claim has been contradicted by Pittner (1991) and Vogel (2001, 2002). Vogel reports the observation of a split among German speakers. They can be divided into three different groups of speakers. The variants are called German A, B, and C. German A is the most liberal and most frequent one, German C the most strict and least frequent. German C is a matching variant, no FRs are possible, if the two cases conflict.

The difference between German A and B can only be seen with one particular conflict, namely, where `m-case` is accusative and `r-case` is nominative. Many German speakers accept both (6-a,b):

(6) a. Ich lade ein wer mir begegnet
I invite(+ACC) who-NOM me-DAT meets(+NOM)

b. Ich lade ein wem ich begegne
I invite(+ACC) who-DAT I-NOM meet(+DAT)

But there is a not too small minority that rejects (6-a). Only very rarely, one can find speakers who even reject (6-b). Pittner (1991) describes the variant that Vogel calls 'German B' (those who do not accept (6-a)) as a variant that allows for FRs if the suppressed case is not higher than the realised case on the

following case hierarchy:

(7) Case hierarchy for German B: (following Pittner, 1991)
nominative $\prec$ accusative $\prec$ oblique (dative, genitive, PP)

German A is 'blind' for the difference between the two structural cases nominative and accusative. For the purpose of our discussion, we might assume the following three constraints (cf., a.o., Vogel, 2002):

(8) **Realise Case**(RC): An assigned case requires a morphological instantiation. (can only be fulfilled by matching FRs)

Realise Case (relativised)(RCr): An assigned case requires a morphological instantiation of itself or a case that is higher on the case hierarchy. (can also be fulfilled by non-matching German FRs, if `r-case` is higher than `m-case`)

Realise Oblique (RO): Oblique Case must be morphologically realised. (this constraint cannot be violated by German FRs)

The ranking of these constraints in German is:

(9) RO $\gg$ RCr $\gg$ RC

Different rankings of faithfulness now yield the three variants, in the following way:[1]

[1] Further constraints are left out here, which are necessary to exclude the strategies M and RES. See (Vogel, 2002) for the full picture and detailed discussion.

(10) *German A*: RO $\gg$ **F** $\gg$ RCr $\gg$ RC
German B: RO $\gg$ RCr $\gg$ **F** $\gg$ RC
German C: RO $\gg$ RCr $\gg$ RC $\gg$ **F**

Table 1 illustrates that the three variants differ in the contexts where they allow for FRs. These contexts themselves can be ordered in terms of markedness. The rankings in (10) predict this finding.

	Matching FRs	possible in German A, B, C
$\prec$	Non-matching FRs that suppress a lower case	possible in German A, B
$\prec$	Non-matching FRs that suppress a higher structural case	possible in German A
$\prec$	Non-matching FRs that suppress oblique case	impossible in German

Tab. 1: Markedness scale of FRs with case conflicts and how they relate to the observed variants of German

Language internal variation, according to the preceding discussion, is variation in terms of ‘tolerance’. There are more liberal and more strict speakers. However, this tolerance is not arbitrary. The relative ranking of the markedness constraints is the same for all of these speakers, they only differ in the rank of faithfulness.

In corpora, we expect differences in the relative frequencies that mirror the scale of FR types in table 1. The less marked, the more frequent a FR should be. In particular:

- For all contexts, correlatives should be more frequent than FRs
- Less marked contexts should occur more frequently than more marked ones
- FRs should occur in less marked contexts relatively more frequently than in more marked ones

3 A corpus investigation

We searched the COSMAS-II corpora[2] of the IDS Mannheim for the three animate *wh*-pronouns *wer* (nominative), *wen* (accusative) and *wem* (dative). The total numbers of instances of sentences with these pronouns in the corpus is given in table 2.

Pronoun	*Total*
wer	166.927
wen	6.327
wem	17.522

Tab. 2: Total number of occurences in the COSMAS-II corpus of written language

[2] We used the largest available corpus, a collection of several corpora of written German, first of all newspaper and magazine articles, prose and scientific literature. According to the IDS homepage, the corpus of 'written language' that we used contains 5.160.576 texts.

Note the extraordinary difference between subject and non-subject *wh*-pronouns. This seems to be due to two independent factors which are both met by *wer*: the tendency of *wh*-pronouns to occur clause-initially, and the tendency of clauses to start with the subject.

We then let the COSMAS-II system select random samples of 500 instances of each of the three pronouns. Animate *Wh*-pronouns have three semantically different usages in German, as FR pronoun, as interrogative pronoun, and as indefinite:

(11) a. *Wer* es glaubt, wird selig (FR)
who it believes becomes blessed

b. Interrogative:

(i) Wer glaubt es ? (main clause)
Who believes it

(ii) Ich weiss wer es glaubt (subordinate clause)
I know who it believes

c. Glaubt es wer ? (indefinite)
believes it someone
'Does someone believe it?'

The distribution of these usages for the three pronouns is given in table 3.[3]

Each of these distributions is highly significant: *wer* is predominantly used as FR pronoun ($\chi^2 = 65.92; p < 0.001$), while *wen* ($\chi^2 = 328.07; p < 0.001$) and *wem* ($\chi^2 = 69.95; p < 0.001$) are predominantly used as interrogatives. Indefinite usages are extremely rare in general. This might be due to the fact that this usage is colloquial, and we are investigating a corpus of written German.

[3] We excluded 3 instances of *wer*, 20 of *wen* and 13 of *wem* because of multiple occurence, listing usages, and other similar reasons.

Pronoun	*FR*	*Interrog.*	*Indef.*
wer	339 (68.20%)	158 (31.80%)	0 (0.00%)
wen	41 (8.54%)	437 (91.04%)	2 (0.42%)
wem	150 (30.80%)	334 (68.58%)	3 (0.62%)

Tab. 3: Distribution of three different usages of *wh*-pronouns

The object pronouns *wen* and *wem* occur both as objects of verbs and as objects of prepositions. As we are only interested in the former, not the latter, we have to tear these usages apart. Table 4 lists the distributions that we found in our sample.

Clause type	*wen*		*wem*	
	Obj. of V	Obj. of P	Obj. of V	Obj. of P
FR	39 (95.1%)	2 (4.9%)	140 (93.3%)	10 (6.7%)
Interrogative	293 (67.0%)	144 (33.0%)	168 (50.3%)	166 (49.7%)

Tab. 4: Usage of *wen* and *wem* as object of verb and preposition

While the distributions for FRs are similar, PPs are relatively rare here, the distribution of PP usages differs largely between *wen* and *wem*. However, the correlation is very small (-0.065), and the χ^2 value of 3.257 is slightly below the level of significance ($.1 > p > .05$). Another difference shows up, when we look at the distribution with respect to main and subordinate clauses. Table 5 shows the relevant figures.[4]

[4] For *wen* we had to take 24 instances out, which were in clausal fragments (14 verbal, 10 prepositional object). With *wem*, it was 26 instances (6 verbal, 20 prepositional object).

	wen		*wem*	
clause type	Obj. of V	Obj. of P	Obj. of V	Obj. of P
matrix	166 (82%)	37 (18%)	83	36
subordinate	113 (54%)	97 (46%)	77 (41%)	112 (59%)

Tab. 5: Distribution of interrogative uses of *wen* and *wem*

For *wem*, we find a weak ($r = 0.28$), but significant ($\chi^2 = 6.08$; $p < .05$) correlation between clause type and more frequent case assigner, such that *wem* is preferably object of a verb in matrix clauses. This finding is highly significant ($\chi^2 = 19.36; p < .001$).We find a weak correlation between *wen* as verbal complement ($r = -0.10$) and its occurrence in a matrix clause, which is also statistically significant ($\chi^2 = 4.53; p < .05$).

Table 6 lists the frequencies of FR and CORR versions of clause-initial FRs in case matching and conflicting configurations. Clause-final FRs are not counted in here, because they cannot have a correlative counterpart. The final column in the table indicates the degree to which a found preference for FR or CORR is statistically significant.

We found FRs in clause-initial and in clause-final position. FRs that stand for the subject of the clause prefer clause-initial position, those that stand for an object, clause-final position. This is expected, as these are the default positions for these grammatical functions. Table 7 lists the distributions.

The crucial findings that are displayed in table 6 are the following:

1. Only matching FRs and non-matching FRs replacing nominative have been found.

r-case	m-case	FR	CORR	Significance
NOM	NOM	274 (89.8%)	31 (10.2%)	***
NOM	AKK	0	2	
NOM	DAT	0	5	
NOM	PP	0	2	
Σ(NOM)		274	40	
AKK	NOM	5 (25%)	15 (75%)	*
AKK	AKK	1 (20%)	4 (80%)	
AKK	DAT	0	3	
AKK	PP	0	0	
Σ(AKK)		6	22	
DAT	NOM	33 (34.4%)	63 (65.4%)	*
DAT	AKK	0	0	
DAT	DAT	1 (5.6%)	17 (94.4%)	***
DAT	PP	0	4	
Σ(DAT)		34	84	

Tab. 6: Frequencies of clause-initial argument FR and CORR clauses relative to case configurations

r-case	m-case	initial	final
NOM	NOM	274 (93.5%)	19 (6.5%)
ACC	NOM	5 (83.3%)	1 (16.7%)
ACC	ACC	1 (12.5%)	7 (87.5 %)
DAT	NOM	33 (84.6%)	6 (15.4%)
DAT	DAT	1 (12.5%)	7 (87.5%)

Tab. 7: Syntactic position of FRs

2. For matching subject FRs, FR is preferred over CORR. This contradicts our expectations. CORR should be more frequent under all conditions.

3. Each of the 5 (ACC) + 33 (DAT) = 38 non-matching FRs use strategy R, strategies M and RES do not occur at all.

4. The overall number of FR and CORR for each of the three cases mirrors well-known preferences for the occurrence of cases in first position, NOM is most likely to occur initially, and ACC dislikes that position most.

5. The relative ranking of contexts given in table 8 displays a highly significant difference between the least marked context NOM-NOM and the rest which can be seen in the exceptional *strong preference for FRs*.

For both dative and accusative matching FRs, 7 out of 8 are clause-final, only 1 is clause-initial. These never occur with a resumptive pronoun anyway. If we exclude these, then the picture changes.

Table 8 shows those environments where FRs have been found at all, and to what degree. The context NOM-NOM is the only one that prefers FR over CORR.

r-case	m-case	FR	CORR
NOM	NOM	274 (89.8%)	31 (10.2%)
DAT	NOM	33 (34.4%)	63 (65.6%)
ACC	NOM	5 (25%)	15 (75%)
ACC	ACC	1 (20%)	4 (80%)
DAT	DAT	1 (5.6%)	17 (94.4%)

Tab. 8: Clause-initial FR and CORR in contexts

This is statistically highly significant for all comparisons. For the context DAT-NOM we also find a statistically significant ($\chi^2 = 6.015; .05 > p > .01$) weak correlation ($r = 0.23; .2 < r < .5$) in comparison with the context DAT-DAT such that the latter context is less likely to occur with an FR than the former. No other comparisons are significant.

Why is FR preferred in NOM-NOM? The theory predicts that CORR should be preferred even here. However, the resumptive pronoun appears to be redundant in those cases:

(12) Wer-NOM es weiss, der gewinnt
who it knows the-NOM wins

This redundancy might be related to the fact that the FR, in addition to realising nominative case, is also located in the correct clause-initial position. Hence, there are already two cues that signal that the FR is subject. The resumptive can serve no additional function.

We compared the NOM-NOM FR and CORR instances in their length, and

found a statistically highly significant ($t = 3.8266; p < .001$) weak correlation ($r = .22; .2 < r < .5$) between FR length and choice of CORR: The longer the FR, the more likely it is doubled by a main clause initial resumptive pronoun.

FR	6.02	
KORR	12.04	***

Tab. 9: Mean number of words between FR pronoun and the first word after the FR in NOM-NOM contexts

The preference for FR in the least marked context, NOM-NOM, can be seen as the exception that proves the rule, namely, that markedness is the driving force behind frequency distributions. The resumptive pronoun becomes redundant in those instances where the FR pronoun bears nominative and the clause-initial FR is the subject of the main clause. The grammatical function 'subject', hence the case of the FR, is already signalled by syntactic position.

4 Conclusion

The corpus study mainly confirmed our expectations about the occurrence of FRs. The interesting exception of NOM-NOM contexts is also driven by markedness. However, the study also shows that structures which are highly marked, but still grammatical, like, for instance, FRs where `r-case` is dative and `m-case` accusative, did not show up at all. There is no difference in frequency between such highly marked structures and clearly ungrammatical structures like, e.g.,

FRs following strategy M.[5] This exemplifies one of the limits of this emprical method.

References

Alexiadou, Artemis and Spyridoula Varlokosta (1995). 'The Syntactic and Semantic properties of Free Relatives in Modern Greek.' *ZAS Working Papers in Linguistics* **5**:1–30.

Bresnan, Joan and Jane Grimshaw (1978). 'The Syntax of Free Relatives in English.' *Linguistic Inquiry* **9**:331–391.

Dayal, Veneeta (1996). *Locality in WH Quantification*. Dordrecht: Kluwer.

Groos, Anneke and Henk van Riemsdijk (1981). 'Matching Effects with Free Relatives: a Parameter of Core Grammar.' *In* Belletti, Adriana, Luciana Brandi, and Luigi Rizzi, eds., *Theories of Markedness in Generative Grammar*, pp. 171–216. Pisa: Scuola Normale Superiore di Pisa.

Grosu, Alexander (1994). *Three Studies in Locality and Case*. London/New York: Routledge.

Harbert, Wayne (1983). 'On the Nature of the Matching Parameter.' *The Linguistic Review* **2**:237–284.

Pittner, Karin (1991). 'Freie Relativsätze und die Kasushierarchie.' *In* Feldbusch, Elisabeth, ed., *Neue Fragen der Linguistik*, pp. 341–347. Tübingen: Niemeyer.

[5] That such a difference can be elicited in a grammaticality judgment experiment, is shown by Vogel and Frisch (2003).

Vogel, Ralf (2000). ‘Towards an Optimal Typology of the Free Relative Construction.’ *In* Grosu, Alex, ed., *IATL8. Papers from the 16th Annual Conference and from the Research Workshop of the Israel Science Foundation “The Syntax and Semantics of Relative Clause Constructions”*, pp. 107–119. Israel Association For Theoretical Linguistics, Tel Aviv: Tel Aviv University.

— (2001). ‘Case Conflict in German Free Relative Constructions. An Optimality Theoretic Treatment.’ *In* Müller, Gereon and Wolfgang Sternefeld, eds., *Competition in Syntax*, pp. 341–375. Berlin: Mouton de Gruyter.

— (2002). ‘Free Relative Constructions in OT Syntax.’ *In* Fanselow, Gisbert and Caroline Féry, eds., *Sonderheft Optimality Theory*, Linguistische Berichte. Hamburg: Helmut Buske Verlag.

Vogel, Ralf and Stefan Frisch (2003). ‘The Resolution of Case Conflicts – a Pilot Study.’ *Linguistics in Potsdam* **21**:91–103.