Amir Zeldes
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
amir.zeldes@rz.hu-berlin.de

On the Productivity and Variability of the Slots in German Comparative Correlative Constructions

Abstract

This paper is concerned with the syntactic productivity and semantic function of the comparative slots in the German comparative correlative construction (*je [COMPARATIVE] ... desto [COMPARATIVE]* and variants), i.e. how prone they are to admitting novel forms under different circumstances, and what each slot is used to express in practice. Using large amounts of corpus data and quantitative productivity measures, it will be shown that comparatives in one slot behave differently from those in the other and from comparatives in general, both in terms of the lexemes they exhibit and in terms of their potential for innovation. Qualitatively, the construction is stereotypically employed to express positive or negative evaluation semantics in the *desto* clause, which depend on a spatiotemporal quantity in the *je* clause. Finally, differences are examined between cases exhibiting nominal subjects and verbal predicates in each clause and cases where these do not appear.

1. Introduction

Comparative correlative constructions (henceforth CCs) are sentences correlating two clauses with respect to comparative adjectives appearing in each clause, as in example (1):

(1) [CLAUSE1 *Je schneller Hans rennt*], [CLAUSE2 *desto schneller wird er müde*]
'The faster Hans runs, the faster he gets tired' (adapted from Beck 1997, 234)
(2) [CLAUSE1 *Je früher*], [CLAUSE2 *desto besser*]
'The sooner, the better'

Such sentences have enjoyed considerable attention, especially in the syntactic literature of recent years, yet surprisingly little has been said about the usage of their central variable component: the comparative adjectives in each clause. This article attempts to address this gap for German by asking several questions about the sorts of lexemes that occupy each clause typically: how does clause1 (hence c1) differ from clause2 (hence c2) in its lexical preferences? How free are speakers to innovate with the comparatives they use in c1 and c2? Are there any differences between usage in sentences like (1), with full subjects and predicates in each clause, and shorter sentences like (2) (hence short CCs), which only contain a comparative after each connector? And how do these observations fit into the syntactic and semantic analyses of CCs in the literature to date?

Looking at previous work on CCs, the two most hotly debated topics so far have probably been the status of their constituent clauses as para- or hypotactic, and the question of their semantic compositionality (see McCawley 1988, Culicover & Jackendoff 1999 for the English equivalents, Beck 1997 for German and den Dikken 2005 for a cross-linguistic account). On the one hand, the fact that many languages use

symmetric forms to realize both clauses has been perceived as a syntax-semantics mismatch (see Culicover & Jackendoff 1999), since the different syntactic function of main and subordinate clauses is expected to be reflected in the forms chosen to represent them (e.g. different connectors, as is the case in German).[1] On the other hand, it is not entirely clear how the special semantics of the correlation between the two comparatives can be derived compositionally from the two clauses, especially if these both look alike, but have different semantic interpretations.[2] In (1), c1 can be interpreted as a sort of conditional to c2, i.e.: *if and when Hans runs faster, he becomes tired that much more quickly*, yet at the same time it does not hold that *Hans runs that much faster, if and when he gets tired more quickly*.[3] For German the situation is somewhat simpler since the corresponding structure is asymmetric already on the surface, namely using the conjunction *je* for the subordinate clause c1 and *desto* for the main clause c2, with the typical verb-second word order in the latter and verb last in the former, as is usual for main and subordinate clauses in German.[4]

Despite this, it is not usually assumed that there is any significant difference between the comparative slot in c1 and c2. The syntactic description of the phenomenon as found in Beck (1997) is of a symmetric CP dominating two externally undifferentiated CPs, each dominating a phrase DegP in their specifier (see Figure 1). Quite independently of the question regarding the analysis of the CPs, I will be concerned with the (a)symmetry of DegP, which is obligatory in all CC clauses (C' may be omitted on both sides in short CCs, or just on one side – i.e. short c1 or c2).[5] Although both DegPs

---

[1] Among the symmetrical languages, the prominence of English with *the* in both clauses has played a role in leading research on CCs in this direction, e.g. in the following example from the British National Corpus (http://www.natcorp.ox.ac.uk/), the structure of one clause mirrors the other:

[$_{\text{CLAUSE1}}$ *The nearer it gets*] [$_{\text{CLAUSE2}}$ *the more worried I become*] (BNC, document A4P)

The clauses may appear paratactic and symmetrical on the surface, but semantics suggest the first clause is in fact subordinate. See also Abeillé et al. (2006) for a discussion of symmetricity in French vs. Spanish CCs.

[2] So much so that CCs have often been used to illustrate Construction Grammar approaches as an example of a construction which requires a partially specified entry in the mental lexicon or 'constructicon', e.g. in Goldberg (2006, p. 5).

[3] This conditional reading has led to the occasional use of the name *comparative conditional* for these constructions. Put more formally, the relationship between the two comparatives is a unilaterally monotonous dependency, though incidentally not necessarily a proportional one. Simplifying somewhat, this corresponds to a formal structure:

$$\forall x, y[g(x) > g(y) \rightarrow f(x) > f(y)]$$

where *g* and *f* are the comparatives modifying their respective CC clauses (see Beck 1997, p. 259).

[4] There are of course both diachronic evidence and synchronic traces of symmetric constructions in German with both *desto … desto* and even *je … je*, but even if these are considered standard, the word order clearly distinguishes main from subordinate clause, e.g.: *Desto lauter sie sind, desto weniger werden sie selbst etwas auf die Beine stellen* 'The louder they are, the less they will get something up on its feet by themselves' (DeWaC, pos. 145401225; see Section 3.1 for information on this corpus).

[5] Short CCs are sometimes considered a case of ellipsis of the subject and predicate, which has been assumed to be a copula verb (e.g. Zifonun et al. 1997: 2338). In fact it is often the case that the introduction of a copula would not make a felicitous sentence, e.g. where a comparative would only fit a telic verb as in *je früher, desto besser* 'the earlier the better'. Here the sense of 'earlier' implies something happening

are realized formally and syntactically in exactly the same way, the data will show that their usage is in fact consistently asymmetric both in the typical filling of the head Deg^0 and in the way this position is used productively with novel items, a mismatch which to my knowledge has received no large scale empirical study to date.

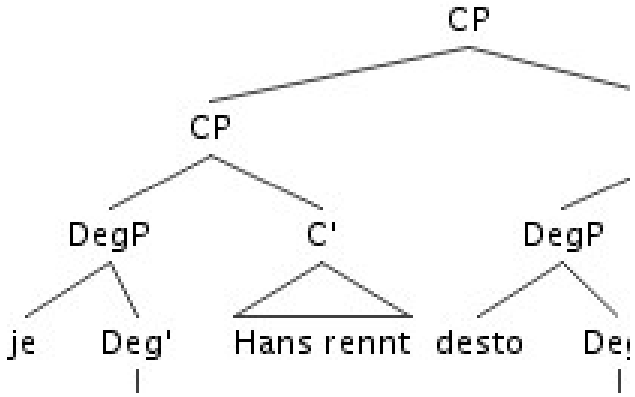

Fig. 1.: Syntactic analysis of German CCs, adapted from Beck (1997, p. 234).

The remainder of the article is structured as follows: section 2 briefly presents approaches to measuring productivity using corpus data. Section 3 presents data from multiple large corpora on the usage of comparatives in and outside of comparative correlatives in German. Section 4 concludes by sketching the asymmetric profile of typical CC usage in German as found in the data, with some suggestions for the interpretation of the differences discussed in section 3.

2. Measuring Productivity

As the scope of this article does not permit an extensive discussion of the possible definitions for the concepts underlying the notion of linguistic productivity, this section will only attempt to give a brief overview of approaches in the empirical paradigm represented in Baayen (1993, 2001, 2009) and related work. In essence, linguistic productivity describes the possibility of forming novel linguistic forms never before heard or produced by a speaker. Productivity is seen as a property of linguistic processes (thus it is a morphological word formation process, or a syntactic process filling an argument structure which are productive, and not affixes, verbs or other words, cf. Bauer 2001, pp. 12-15). A prerequisite for quantitative empirical work on productivity is the view that productivity is a gradual property, and not binary or even categorical. Thus there is no dichotomy between rules of grammar which are productive and those which are not (e.g. past tense with *-ed* vs. vowel changes for weak and strong verbs in English), but rather some processes may be extremely unproductive (and indeed, novel strong past tense forms do occur; for discussion see e.g. Clahsen 1999, Ullman 1999, McClelland & Patterson 2002). A distinction between 'productive', 'unproductive' and 'semi-productive' (as made by some researchers, see Bauer 2001, 15-20) is also unhelpful in this context, both intuitively, since some processes are perceived to be more productive

---

rather than a continuous state (i.e. 'the sooner it happens' or the 'the sooner you do it', but probably not 'the sooner it is'). In any case as we will see in Section 3, short CCs actually behave quite differently from other CCs in their preferences for certain lexemes and in their producitivity.

than others on a scale, (e.g. Dutch *ver-* vs. *-ster* in Baayen 2009, pp. 904-907, or English deadjectival nominalization in *-ness*, *-ity* or *-cy*, cf. Plag 1999, C. 2; see also Bauer 2001, C. 1-2) and in practice, since data-based measurements lend themselves to normalized scales.

The criteria for a productive formation in most work tend to focus on novelty, regularity and transparency (see Bauer 2001, pp. 34-58). That is to say, a process is productive if and only if it produces forms never before generated or received by the speaker, which result regularly from the process and the components on which it operates, and the resulting forms can be understood with predictable meaning in a fashion consistent with other formations from the same process. However in reality it is impossible to directly or reliably evaluate the novelty, transparency and regularity indicative of productivity for all items associated with a process (even for one speaker it is impractical to establish whether or not she or he have produced or received a particular formation in the past, let alone for the linguistic community as a system, as Bauer (2001, pp. 34-35) points out). Baayen (2001, 2009) therefore suggests that different aspects of productivity can be assessed, at least for a certain register, from corpus data, using the type and token frequencies of a word formation, as well as the frequency of items appearing only once in the corpus (*hapax legomena*), which are assumed to be a superset of the neologisms therein. In particular, Baayen suggests using three different measures:

p1: Extent of Use = $V(C,N) = \frac{Types_C}{N}$

p2: Hapax-conditioned Degree of Productivity = $\frac{V(1,C,N)}{V(1,N)} = \frac{Hapax_C}{Hapax_N}$

p3: Category-conditioned Degree of Productivity = $\frac{V(1,C,N)}{N(C)} = \frac{Hapax_C}{Tokens_C}$

Although it is clearly not the case that there is a constant ratio between hapax legomena and 'true' neologisms in every corpus, these measures often seem to correspond to linguists' intuitions about productivity. p1 simply specifies how large a vocabulary the process has produced in *N* tokens of data, p2 the proportion of unique items in the corpus coming from the process and p3 the proportion of unique items within tokens belonging to the process.[6]

To illustrate the use of these measures, I use a corpus of 5 years of the German computer magazine "c't Magazin" (CT, 1998-2002, some 15 million tokens), comparing data for three adjective forming suffixes with different degrees of productivity: *-bar*, *-lich* and *-sam*, which form such adjectives as *lesbar* 'readable' (from *lesen* 'to read'), *freundlich* 'friendly' (from *Freund* 'friend') and *einsam* 'lonesome' (from *ein* 'one'). The results for these suffixes are summarized in Table 1.

[6] In Baayen's notation V(C,N) stands for the vocabulary size of a morphological category C in N tokens, or in the context discussed here, the normalized type count of the output of a process. V(1,N) is the amount of hapax legomena in the corpus (vocabulary types with frequency = 1, or simply $Hapax_N$), and V(1,C,N) is the amount of types from the relevant category C with a frequency of 1 (or $Hapax_C$). N(C) is the token count for all occurrences of the category in the data.

	-lich	*-bar*	*-sam*
Tokens	59472	26865	7691
Types	1222	896	74
Hapax	483	354	24
p1	0.002054	0.000061	0.000005
p2	0.001356	0.000994	0.000067
p3	0.008121	0.013176	0.003120

Tab. 1: Productivity measures for *-lich*, *-bar* and *-sam* in the CT corpus

As the table shows, the *-sam* formation is the least productive, with few hapax legomena and the lowest score on all three measures. *-lich* exhibits a larger vocabulary than *-bar*, indicating it has been more productive in the past, but *-bar* has a higher proportion of hapax legomena and consequently higher p2 and p3, indicating it is now easier to form novel adjectives with this suffix. This probably corresponds with most intuitive judgments (essentially the same results with a different corpus may be found in Evert & Lüdeling 2001), as *-sam* forms virtually no new forms in present-day German, and *-lich* is more restricted than *-bar*, which can form adjectives expressing potentiality from almost any transitive verb stem.

However applying the measures to different sample sizes from each process leads to skewed results: the more words we have examined from a certain category, the more likely it becomes that the next word will not be novel (since we already 'know' more words). It is therefore necessary to compare the measures at a fixed sample size (e.g. *n* thousand samples from each process, all in the same corpus, see also Gaeta & Ricca 2006), which also allows statistical significance to be computed. The linguistic interpretation of the different measures can best be illustrated by plotting the development of vocabulary size across the corpus. This is achieved using vocabulary growth curves (VGCs, see Evert 2004), which plot the amount of tokens on the x-axis and the amount of types at that point on the y-axis. Thus each newly encountered hapax legomenon raises the curve, but as more and more familiar items are encountered, it becomes gradually flat, showing that the process is approaching saturation in the data.

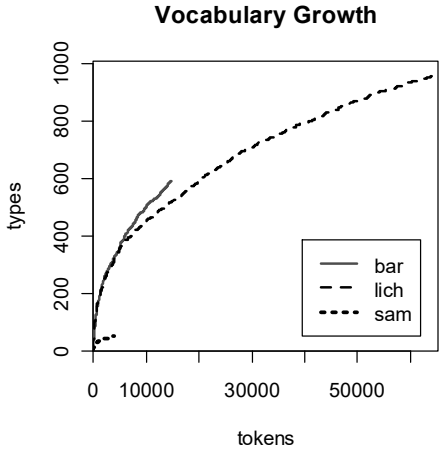

Fig. 2: VGCs for *-lich*, *-bar* and *-sam* in the CT corpus.

Figure 2 illustrates the higher productivity of *-bar*, which has a shorter curve (fewer types in the data, lower p1) but a steeper incline (higher proportion of hapax

legomena, thus higher p3) along the curve. The *-sam* curve is very short as items with this suffix are rare, but already much flatter than the other two curves – even at this small sample size, the large majority of types from this process have already been seen. Since longer curves offer more chances for different types to occur, but with a progressively lower chance for novel hapax legomena, significance can only be evaluated based on the smallest sample size. The following section describes the corpora used in this study in more depth, followed by an analysis of differences in the productivity of the comparative formation in c1 and c2 and a study of lexical preferences for each slot.

3. The Data for *je … desto*

3.1. Corpora

Given the relative infrequency of CCs, a large corpus (or several corpora) is required in order to find a sufficiently large database of clauses with *je* and *desto*. Thus although an examination of a variety of genres would be desirable in principle, the main available choices of written language with sufficient size are newspaper language and Web data. I will therefore use the controlled CT corpus mentioned above, and the largest available web corpus of German, the uncontrolled DeWaC corpus (Baroni & Kilgarriff 2006, approx. 1.7 billion tokens). In order to admit some information on usage of the construction in the spoken medium, I also use German Parliament Proceedings (GPP, from 1996 to February 2003, totaling some 37 million tokens). As it turns out, however, the construction is rather rare in the proceedings register (about half as frequent compared to the CT corpus). For this reason a further 27 million tokens were taken from the German version of the proceedings of the European Parliament (Europarl, Koehn 2005; partly original German sentences, partly translated from 10 other European languages), producing around the same size dataset for both genres. The use of translated language in this context is not optimal (though arguably expert translations forming German sentences are a valid genre in and of themselves; for a discussion see Olohan 2004, C. 3 and 7), yet data from Europarl actually matches distributions in the proceedings of the German Parliament surprisingly closely.[7]

With tokenized and part-of-speech tagged data at hand, frequencies were extracted for all predicative/adverbial adjectives in the corpora ending in the comparative ending *-er*,[8] and the resulting 3.5 thousand lexemes were manually sorted for plausibility as a comparative adjective (filtering out both wrongly tagged non-adjectives such as *eBay-Webserver* and non-comparative, attributive cases such as *genannter* 'named'). For the remaining 2,000 or so comparative lexemes, total frequencies (potentially including wrongly tagged attributive cases) and frequencies after *je* and *desto* were extracted, as well as frequencies in the sequence *je [COMPARATIVE] , desto [COMPARATIVE]*, where the comma was optional. Using the methods introduced in section 2, it is possible to calculate

[7] For instance the top 10 lexemes for comparatives after *je* match 9 out of 10 in the two corpora, with *früher* 'earlier' replacing *später* 'later' in Europarl, and 7 out of 10 after *desto*. At the same time the CT data is less conformant with both proceedings corpora in lexical choices than the latter are between themselves.

[8] The corpora were tagged using the freely available TreeTagger (Schmid 1994) and searched with the Corpus Workbench (CWB, Christ 1994) for the STTS part-of-speech tag ADJD (see Schiller et al. 1999 for the tagset).

the productivity measures for the comparative formation in the available corpora, excluding those cases which follow *je* or *desto*. These results are presented in Table 2.

	CT	Europarl	GPP	All Corpora
corpus tokens	14596537	27317723	36723139	78637399
corpus types	595022	283389	443949	1010539
corpus hapax leg.	356075	140730	222221	565020
comparative tokens	30548	41857	49866	122271
comparative types	1149	1160	1383	1969
comparative hapax l.	515	494	648	776
p1	0.00007872	0.00004246	0.00003766	0.00002504
p2	0.00144632	0.00351027	0.00291602	0.00137340
p3	0.01685871	0.01180209	0.01299483	0.00634656

Tab. 2: Corpus statistics and productivity measures for comparatives outside CCs

A direct comparison between the data for each corpus should be avoided, since they are of different sizes and thus have different chances of realizing fewer or more types and hapax legomena. However it should become clear from the vocabulary and hapax counts that CT is the richest corpus, with more types and hapax legomena than the other two corpora, despite having the least amount of tokens. This is understandable, as the magazine contains a variety of text types (reviews, editorials, readers' letters) and a high amount of unique technical terms increasing both vocabulary and neologisms. For the comparative counts the situation is more moderate, but CT still has the highest type/token ratio and almost as many types as the other corpora, thus revealing again the largest variety for the smallest corpus.

3.2. Differences in Productivity for c1 and c2

Applying the measures presented in section 2 to the slots c1 and c2 reveals differences in their productive potential to manifest new items as predicted by ratios of hapax legomena. Following Barðdal (2006), who examines the productivity of ditransitive verbs in Icelandic and Kiss (2007), who applies Baayen's measures to the nominal slot of PPs with determinerless singular nouns in German, I will treat the filling of the comparative position in each CC clause as a productive process, with the choice of comparative paralleling the choice of a stem in a morphological process such as affixation. Figure 3a plots the vocabulary growth curve for comparatives in c1 and c2 in all corpora as compared with 3000 randomly selected comparatives outside of CCs equally distributed between all three corpora.

It is immediately clear that non CC comparatives (the top curve, *comp*) are more productive than CC comparatives (significant test of equal proportions at $p<.01$ for an equal sample size). Since the CC sample only covers less than 1500 tokens, the data is extrapolated to show expected development of the curves using a finite Zipf-Mandelbrot model (FZM, see Evert 2004), which provides a good estimate of the expected divergence of the curves given more data from the same register. Results also show the c2 curve (*c2* or the extrapolation *fzm2*) to be significantly ($p<.01$) more productive than the c1 curve (*c1* or *fzm1*). This relationship is also true for each genre separately, though splitting the corpus would result in figures rather small for a productivity study and insufficient for a significant result.

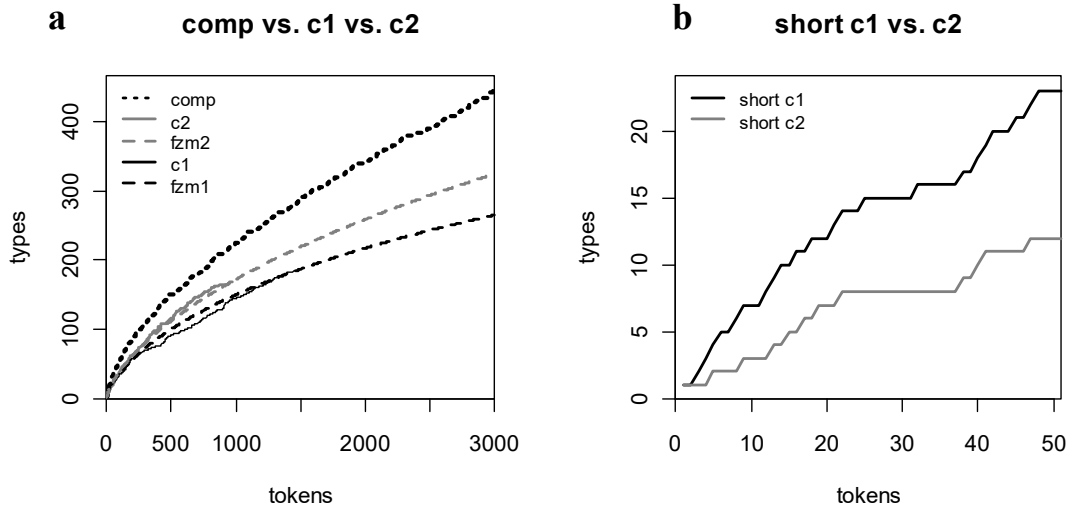

Fig. 3: On the left (a), comparatives outside CCs (top curve) are more productive than c2 (middle), which is more productive than c1 (bottom). Dashed curves predict the further development of CC vocabulary based on a finite Zipf-Mandelbrot model (FZM). On the right (b), a very small sample suggests short c1 is more productive than short c2.

Thus results show c2 is more open to lexical variation than c1. Another interesting question already raised in section 1 is whether short CCs behave in the same way as other CCs. Surprisingly, the data exhibits a trend in the opposite direction (Figure 3b), though numbers are too small to be significant. The lexeme responsible for this situation is largely *besser* 'better', which forms approx. 73% of the data, or 37 matches for short c2. Since a much larger sample is needed in order to establish a meaningful trend, the experiment is repeated with DeWaC. Though uncontrolled and therefore likely more heterogeneous and possibly more productive, this dataset has the advantage of containing over 1800 short CCs (showing the rarity of this construction: only about .0088 times per 10,000 tokens, or less than one in a million!). Results repeat the same pattern (Figure 4).

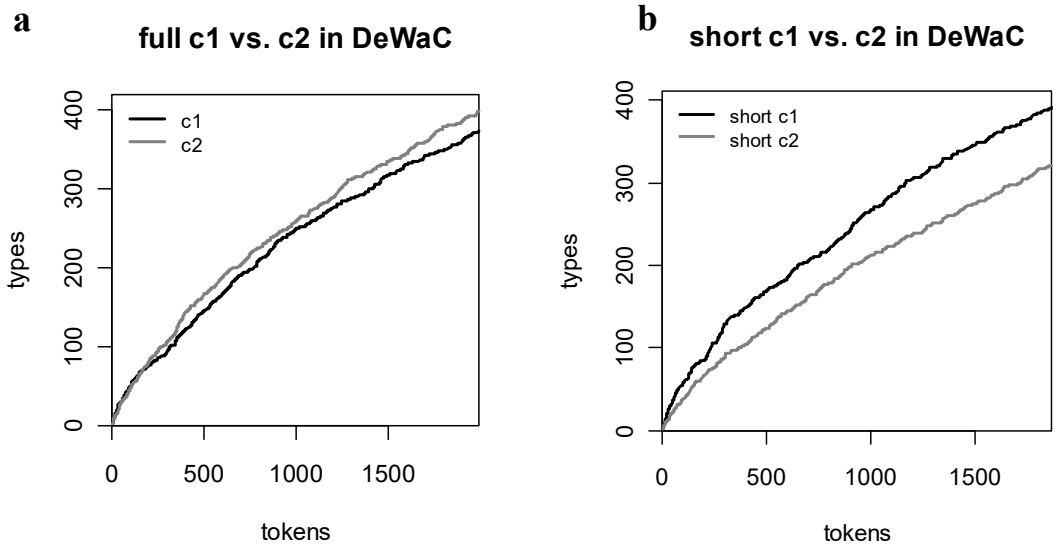

Fig. 4: VGCs for same sized samples of short and long c1 and c2 in DeWaC.

c2 is only a little more productive than c1 in full CCs (p3=.028 for c1 and .031 for c2), though significantly thanks to the large sample (Figure 4a). At the same time it is considerably less productive in short CCs (see Figure 4b), in which *besser* 'better' comprises an overwhelming majority of 55% of c2 cases (1022 matches, leading to a p3 score of .13 for c1 vs. .11 for c2 and a noticeably flatter VGC). From this influence of *besser* it should become clear that specific lexical items can be an important factor in explaining the differences in the behavior of the two slots. With these results at hand, the next section therefore turns to examine the lexemes occupying each slot in more detail.

3.3. Lexical Preferences for c1 and c2

Though it is impractical to examine each and every attested CC in the corpus, cross-sections of lexical behavior in each slot from very frequent, moderately frequent and rare items can reveal some trends. Table 3 shows the frequencies of different comparative adjectives in each corpus and in all corpora put together in c1 and c2, as well as the total frequency for each comparative in general. Trends which are systematic and register independent should appear in all corpora, whereas mixed results might lead to doubts as to any meaningful patterns in the data.

	All Corpora			**CT**		**GPP**		**Europarl**	
type	**freq**	**c1**	**c2**	**c1**	**c2**	**c1**	**c2**	**c1**	**c2**
besser	18270	70	212	19	79	32	46	19	87
später	7983	22	5	4	1	15	2	3	2
stärker	6844	65	56	15	26	21	13	29	17
ferner	5975	0	0	0	0	0	0	0	0
länger	4659	179	23	44	17	93	4	42	2
höher	3330	179	76	109	43	55	26	27	7
größer	3126	195	120	117	39	41	39	37	42
…									
schwieriger	1344	7	24	4	9	1	6	2	9
kleiner	878	79	11	54	8	13	3	12	0
…									
positiver	116	0	3	0	1	0	1	0	1
dunkler	112	13	4	12	4	0	0	1	0
wahrscheinlicher	103	0	10	0	6	0	0	0	4
…									
lockerer	25	3	0	3	0	0	0	0	0
wärmer	25	1	0	0	0	0	0	1	0
mühsamer	24	0	1	0	1	0	0	0	0

Tab. 3: Frequencies for comparatives in and outside of CCs in each corpus (see text for translations).

An examination of the table reveals some very strong lexical preferences which are remarkably consistent across the corpora. The generally most frequent comparative, *besser* 'better', is also the most frequent lexeme in c2 in all corpora. It is not, however, the most frequent in c1 – *besser* is three times less frequent in this position in total, outranked at a large margin by otherwise considerably less frequent lexemes such as *länger* 'longer' or *höher* 'higher'. These lexemes are in turn much more frequent in c1

than in c2, thus exhibiting the opposite asymmetry. The fourth most frequent comparative, *ferner* 'further' is not used in CCs at all, though this is unsurprising since it is almost exclusively used as a lexicalized adverb with the sense 'furthermore'.[9] Some lexemes are much more balanced, such as *stärker* 'stronger', or have a less overwhelming imbalance, such as *größer* 'bigger'. Turning to mid-frequent CC comparatives, we find consistent asymmetries yet again, where all three corpora show the same preference of some comparatives for either c1 (e.g. *kleiner* 'smaller') or c2 (e.g. *schwieriger* 'more difficult'). Finally items that are rare or even hapax legomena in each corpus, potentially indicating less entrenched, productively formed CCs that had not been produced by the speaker/writer before (see section 2), also cluster around slots: *dunkler* 'darker', *lockerer* 'looser' and *wärmer* 'warmer' appear mostly or exclusively in c1 and *positiver* 'more positive', *wahrscheinlicher* 'more likely' and *mühsamer* 'more laborious' prefer c2.

How can these results be interpreted? A lexicalization of large lists of lexemes to prefer one position or the other seems unlikely, especially considering the evident preferences of rather infrequent items across corpora (working under the assumption that at least moderate frequency is a prerequisite for lexicalization). A more careful look at the senses of the adjectives reveals a likelier semantic explanation: c1 prefers spatiotemporal conditions, whereas c2 provides an evaluation which typically passes a subjective judgment on the favorability or likelihood associated with the increase of the condition in c1. This interpretation is evident simply by composing sentences from the most frequent c1s and c2s:

(3) *Je höher, desto besser* 'the higher the better'
(4) *Je länger, desto schwieriger* 'the longer the more difficult'

Such sentences also form the typical cases of short CCs (see below). Cases which appear semantically more spatiotemporal but still appear in c2, such as *größer*, with its more balanced profile, merit a closer look. A qualitative examination of c2 sentences of this sort often reveals that such lexemes may assume a rather neutral role when used to modify a subject noun, which in turn supplies the evaluative semantics. This may appear in c2 (example 5), but also in c1 (6):

(5) *je länger man den Rechner laufen lässt […] desto größer die Gefahr , dass sich der Schaden noch vergrößert*
'The longer the computer is allowed to run […] the greater the risk that the damage will increase even more'
(CT 2000 vol. 6 p. 116 segment title "Praxis: Datenrettung per Diskeditor")
(6) *Je größer der Abstand zur Vollaussteuerung […] , desto besser*
'The greater the distance to complete amplification […] the better'
(CT 1998 vol. 1 p. 102 segment title "Prüfstand: Soundkarten")

In (5), the c1 spatiotemporal condition 'time running' is correlated with the idea of 'risk', however rather than formulating the notion adjectivally (*desto gefährlicher* 'the riskier'),

[9] A true comparative sense is still possible nonetheless, e.g.: *Nichts lag aber der DDR-Diktatur ferner als der Frieden* 'But nothing was further removed from the GDR dictatorship than peace' (GPP, July 6 2000, session 114); such cases are however quite rare and unattested in CCs.

größer ‘greater’ is used to modify the ‘risk’. Though ‘greater’ basically refers to a measurable expanse (thus also spatiotemporal in an extended sense), the reading as a whole is still evaluative (risky and hence negative). Similarly in (6) *größer* does not specify the spatiotemporal semantics by itself but rather qualifies *Abstand* ‘distance’ (which could have also been specified with a single comparative, e.g. *weiter* ‘farther’). The opposite situation, where an apparent evaluative qualifies c1, is less frequent and also turns out to blend into a larger spatiotemporal condition in most cases, as in (7):

(7) *Je besser die Komprimierung ist, um so höher fällt ohne zusätzliche Speichererweiterung die nutzbare Auflösung für größere Grafiken aus*
‘The better the compression is, the higher the usable resolution turns out for bigger graphics without additional memory expansion’
(CT 1999 vol. 1 p. 116 segment title “Prüfstand: Laserdrucker”)

Although *besser* appears in c1 unusually, it qualifies a rate of compression (a sort of spatiotemporal condition), and this is in turn evaluated in terms of better print resolution. In either case it seems that adjectives with a less specific semantic content can be used to modify CC subjects as a sort of ‘light comparatives’, where the subject noun of the modified clause specifies the typical meaning supplied by c1 or c2. In these cases, there is therefore still a tendency for c1 to contain a spatiotemporal condition, and c2 a dependent evaluative.

Moving on to the short CCs, a more extreme set of preferences can be observed by comparison. Table 4 shows frequencies in CCs in total vs. short CCs for each lexeme.

type	freq	c1 total	c2 total	c1 short	c2 short
besser	18270	70	212	0	37
länger	4659	179	23	2	0
schneller	4423	88	40	8	0
höher	3330	179	76	0	1
größer	3126	195	120	4	0
schlechter	1665	10	16	0	1
früher	1483	15	0	11	0
schlimmer	1435	2	3	1	0
deutlicher	1338	5	12	1	0
billiger	1219	2	4	1	0
häufiger	1106	11	13	0	1
kleiner	878	79	11	2	0
sicherer	524	5	7	0	1
kürzer	383	26	6	3	0
…					
ergonomischer	8	0	1	0	1
frecher	7	1	1	0	1
hilfloser	4	0	1	0	1
teurer	4	0	0	2	2
unsolider	2	1	0	1	0
reißerischer	1	1	0	1	0

Tab. 4: Preferences for short CCs.

The data shows a stronger bias for short CCs, where the most frequent comparative *besser* is not only strongly preferred in c2, but does not occur at all in c1 (a ratio of 37 to zero, thus clauses of the type *je besser, desto [COMPARATIVE]* are entirely unattested). Conversely, the most common lexeme in c1 is *früher* 'earlier' (11 times in c1 but 0 in c2), closely followed by *schneller* 'faster' (8 and 0 respectively). Figure 5 illustrates the distribution of lexemes in each short slot.

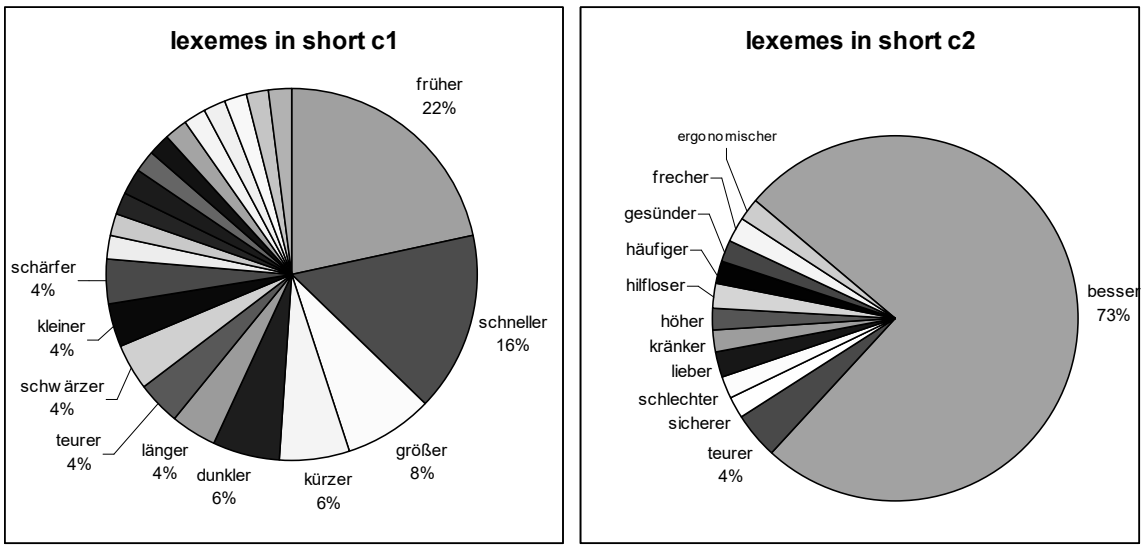

Fig. 5: comparative lexemes in short c1 and short c2

The pattern is markedly different on each side. In fact, only one item appears in both slots: *teurer* 'more expensive'. While an evaluative reading of 'expensive' = 'bad' requires little explanation, an extension of its sense to a spatiotemporal reading requires an explanation. In fact, both occurrences of this comparative in c1 are subsequently evaluated in c2, namely:

(8) *je teurer, desto besser* 'the more expensive the better'
(CT 2002 vol. 9 p. 6 segment title "Inhalt: 09/2002")
(9) *Je teurer, desto schlechter* 'The more expensive the worse'
(CT 1998 vol. 22 p. 170 segment title "Software: Übersetzung")

where (9) is used sarcastically in reference to expensive machine translation software, and (8) is used negatively to say that this rule does not apply to printer inks and paper types. Although *teurer* in these latter cases is not spatiotemporal in any but the most transferred sense, the structure of relating a quantity to an evaluation still bears some resemblance to other cases.

A possible explanation for the more pronounced tendencies in short CCs is that the typical semantics of each clause can only be expressed in the comparative itself, whereas long CCs may distribute the sense of each correlate between the comparative and the subject phrase or even predicate (if both are supplied). This means that ambivalent lexemes like *größer* are read as spatiotemporal by default (notwithstanding additional meanings supplied by context and not realized overtly). At the same time the

overwhelming dominance of *besser* in c2 seems to suggest the choice of a short CC is most appropriate for comparatively simple evaluations, which fits well with the fact that this slot is also the least productive. This is not to say that productivity is ruled out in c2 – only that it is less likely, much like in the case of less productive morphological affixes.

4. Summary - A Profile of German CCs in Use

The picture of CC usage in German arising from the data used in this study is of a semantically asymmetric construction, correlating a scalar, usually spatiotemporal quantity in c1 with an evaluation of the effect of a change in this quantity in c2. In c1, typical spatial examples are size and distance, such as 'bigger' or 'farther', and typical temporals are either a flexible point in time, especially using the notion of 'earlier', or durations such as 'longer' (the latter can also function spatially for distance of course). Some extended senses also found frequently are color terms (e.g. 'darker', 'brighter' or even actual colors like 'blacker', 'whiter'), where perhaps depth of color is meant, as an extension of spatial depth, and references to price as in 'more expensive' or 'cheaper' (though the prevalence of this category may be connected to the rather economically oriented genres examined). For this last case, a true spatiotemporal interpretation is not obvious, though it is clear why such a quantity is often correlated with an evaluation of advantageousness (this is coded in the opposite c1-c2 order in the expression *value for money*, and in the canonical CC order in the German *Preisleistungsverhältnis* 'price-benefit-ratio'). The c1 slot is overall less productive than c2, meaning novel spatiotemporals arise somewhat less frequently.

In c2 we find both direct evaluations of quality, notably 'better' (or less frequently 'worse'), but also often evaluations of probabilities – 'more likely', 'riskier', 'more certain', and more semantically specified evaluations such as 'healthier', 'more difficult', 'more laborious'. We also find some (though fewer) spatiotemporals, notably 'greater' or 'bigger', especially when these function as a sort of semantically underspecified 'light comparatives', qualifying a noun supplying the evaluative meaning. Thus we get *desto größer die Wahrscheinlichkeit* 'the greater the probability' instead of *desto wahrscheinlicher* 'the more likely', or *desto größer die Gefahr* 'the greater the danger' instead of *desto gefährlicher* 'the more dangerous'. This slot is also significantly less productive than comparatives outside of CCs, but more productive than c1, meaning novel ways of evaluating c1 conditions are more likely than such novel spatiotemporal circumstances in full CCs.

The examination of short CCs has shown them to adhere even more closely to the lexical stereotypes, possibly since there is no more possible recourse to the semantics of other phrases (subject, predicate or other adverbials) to supply the spatiotemporal or evaluative meaning. At the same time they are also the least productive, but with the opposite internal relationship: c2 is much less productive than c1, with *besser* filling a sweeping majority of short c2s. This implies that this construction tends to be chosen precisely in cases where the message of the utterance is simply a positive judgment on some condition, leaving more variety in the expression of the condition itself; if further nuances of the evaluation are required (e.g. it is better in the sense of 'more certain', 'healthier', etc.), the short CC is apparently less preferred. Still, items other than *besser* are clearly possible, and *besser* also occurs in long CCs in c1, and then often as a 'light comparative' in much the same way as 'bigger' or 'greater'.

The theoretical status of the observations made here is not yet clear. On the one hand, it is unquestionably true that: 1. lexically, many comparatives can and do appear in both c1 and c2 which do not obey the prototypical semantics portrayed here, and 2. productively, both slots are capable of hosting novel comparatives presumably not heard before by the speaker. On the other hand, multiple, rather large datasets have shown that the properties charted here for each slot show significant and consistent differences in the propensity for innovation and an unequivocal preference for certain lexemes and types of lexemes. These facts require an explanation, as they seem to suggest speakers have implicit knowledge of how to use CCs, which must be stored somehow in reference to the meaning of the construction as a whole (this brings to mind the 'constructicon' account of Construction Grammar mentioned in section 1, as in Goldberg 1995, 2006). Accounting for such facts of usage becomes even more important if we view the emergence of grammar as a gradual codification of such 'soft constraints', which can be more or less categorical (cf. Bresnan et al. 2001; the soft constraints of one language, or even language stage, may be mirrored in the categorical constraints of another; see also the articles in Bybee & Hopper 2001).

Facts of usage not touched upon here but meriting further study are the interaction between the choice of c1 and c2 (particularly likely pairs and conditional probabilities in each direction), both semantically and through preferred lexicalized orders. It is conceivable that certain frequent CCs, especially where the correlation is bilateral, form steady c1-c2 pairs in a similar way to so-called irreversible binomials (like English *black and white* but usually not *#white and black*; see Malkiel 1959; Müller 1997; Ross 1980). The difference between full CC clauses with subject and verb and those with a subject but no VP also requires a separate investigation, as well as the behavior of cases where only one clause is short. Finally, a cross-linguistic analysis to examine whether the trends in German CC data are mirrored in CCs in other languages is important for establishing whether these results reveal general semantic factors (the typical use of comparatives language-independently, or constraints imposed by world knowledge) or rather language specific preferences.[10] The study of the semantics of CCs is therefore far from complete, and offers a rich environment for comparing the use of what seems like a single category, comparatives, but turns out to be very differentiated depending on its embedding context.

References

Abeillé, A., Borsley, R.D. & Espinal, M.T. (2006): The Syntax of Comparative Correlatives in French and Spanish. In: Müller, S. (ed.), *Proceedings of the HPSG06 Conference*. CSLI Publications.

Baayen, R. H. (1993): On Frequency, Transparency, and Productivity. In: Booij, G.E., van Marle, J. (eds.), *Yearbook of Morphology 1992*. Dordrecht: Kluwer, 181-208.

Baayen, R. H. (2001): *Word Frequency Distributions*. (Text, Speech and Language Technologies 18.) Dordrecht / Boston / London: Kluwer Academic Publishers.

Baayen, R. H. (2009): Corpus Linguistics in Morphology: Morphological Productivity. In: Lüdeling, A. & Kytö, M. (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 899-919.

[10] Preliminary results from a study of similar data in English but with different goals (Zeldes 2009) suggest that spatiotemporal and evaluative semantics are indeed typical for English CCs as well, though the particulars for some lexemes and the degrees of productivity behave in a noticeably different manner.

Barðdal, J. (2006): Predicting the Productivity of Argument Structure Constructions. In: *The 32nd Annual Meeting of the Berkeley Linguistics Society*. Available at: http://ling.uib.no/barddal/BLS-32.barddal.pdf.
Baroni, M. & Kilgarriff, A. (2006): Large Linguistically-processed Web Corpora for Multiple Languages. In: *Proceedings of EACL 2006*, Trento, Italy, 87-90.
Bauer, L. (2001): *Morphological Productivity*. (Cambridge Studies in Linguistics 95.) Cambridge: Cambridge University Press.
Beck, S. (1997): On the Semantics of Comparative Conditionals. *Linguistics and Philosophy* 20, 229-271.
Bresnan, J., Dingare, S. & Manning, C.D. (2001): Soft Constraints Mirror Hard Constraints: Voice and Person in English and Lummi. In: *Proceedings of the LFG '01 Conference*, Hong Kong. Available at: http://www.stanford.edu/~bresnan/lfg01.pdf.
Bybee, J.L. & Hopper, P. (eds.) (2001): *Frequency and the Emergence of Linguistic Structure*. (Typological Studies in Language 45.) Amsterdam: John Benjamins.
Christ., O. (1994): A Modular and Flexible Architecture for an Integrated Corpus Query System. In: *Proceedings of Complex 94*. Budapest, 23-32.
Clahsen, H. (1999): Lexical Entries and Rules of Language: A Multidisciplinary Study of German Inflection. *Behavioral and Brain Sciences* 22, 991-1060.
Culicover, P.W. & Jackendoff, R. (1999): The View from the Periphery: The English Comparative Correlative. *Linguistic Inquiry* 30(4), 543-571.
den Dikken, M. (2005): Comparative Correlatives Comparatively. *Linguistic Inquiry* 36(4), 497-532.
Evert, S. (2004): A simple LNRE model for random character sequences. In: *Proceedings of JADT 2004*, 411-422.
Evert, S. & Lüdeling, A. (2001): Measuring Morphological Productivity: Is Automatic Preprocessing Sufficient? In: Rayson, P., Wilson, A., McEnery, T. Hardie, A. & Khoja, S. (eds.), *Proceedings of Corpus Linguistics 2001*. Lancaster, 167-175.
Gaeta, L. & Ricca, D. (2006): Productivity in Italian Word Formation: A Variable Corpus Approach. *Linguistics* 44(1), 57-89.
Goldberg, A. E. (1995): *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago and London: University of Chicago Press.
Goldberg, A. E. (2006): *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
Kiss, T. (2007): Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. *Zeitschrift für Sprachwissenschaft* 26(2), 317-345.
Koehn, P. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proceedings of the Tenth Machine Translation Summit. Phuket, Thailand*, 79-86.
Malkiel, Y. (1959): Studies in Irreversible Binomials. *Lingua* 8, 113-160.
McCawley, J. D. (1988): The Comparative Conditional in English, German and Chinese. In: *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, 176-187.
McClelland, J.L. & Patterson, K. (2002): Rules or Connections in Past-tense Inflections: What Does the Evidence Rule Out? *Trends in Cognitive Science* 6, 465–472.
Müller, G. (1997): Beschränkungen für Binomialbildung im Deutschen. Ein Beitrag zur Interaktion von Phraseologie und Grammatik. *Zeitschrift für Sprachwissenschaft* 16(1/2), 5-51.

Olohan, M. (2004): *Introducing Corpora in Translation Studies*. London and New York: Routledge.
Plag, I. (1999): *Morphological Productivity: Structural Constraints in English Derivation*. (Topics in English linguistics 28.) Berlin: Mouton de Gruyter.
Ross, J. (1980): Ikonismus in der Phraseologie. *Zeitschrift für Semiotik* 2, 39-56.
Schiller, A., Teufel, S., Stöckert, C. & Thielen, C. (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Technical report, University of Stuttgart and University of Tübingen. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf.
Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK. Available at: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.
Zeldes, A. (2009): Quantifying Constructional Productivity with Unseen Slot Members. In: *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*, June 5, Boulder CO, 47-54.
Zifonun, G., Hoffmann, L. & Strecker, B. (eds.) (1997): *Grammatik der deutschen Sprache, Bd. 3*. (Schriften des Instituts für deutsche Sprache 7.) Berlin / New York: de Gruyter.