Submitted to *Morphology* (JOMO) manuscript No.
(will be inserted by the editor)

The plural interpretability of German linking elements

Roland Schäfer · **Elizabeth Pankratz**

Received: date / Accepted: date

Abstract In this paper, we take a closer theoretical and empirical look at the linking elements in German N1+N2 compounds which are identical to the plural marker of N1 (such as *-er* with umlaut, as in *Häus-er-meer* 'sea of houses'). Various perspectives on the actual extent of plural interpretability of these pluralic linking elements are expressed in the literature. We aim to clarify this question by empirically examining to what extent there may be a relationship between plural form and meaning which informs in which sorts of compounds pluralic linking elements appear. Specifically, we investigate whether pluralic linking elements occur especially frequently in compounds where a plural meaning of the first constituent is induced either externally (through plural inflection of the entire compound) or internally (through a relation between the constituents such that N2 forces N1 to be conceptually plural, as in the example above). The results of a corpus study using the DECOW16A corpus and a split-100 experiment show that in the internal but not external plural meaning conditions, a pluralic linking element is preferred over a non-pluralic one, though there is considerable inter-speaker variability, and limitations imposed by other constraints on linking element distribution also play a role. However, we show the overall tendency that German language users do use pluralic linking elements as cues to the plural interpretation of N1+N2 compounds. Our interpretation does not reference a specific morphological framework. Instead, we view our data as strengthening the general approach of probabilistic morphology.

Keywords compounds · linking elements · German · probabilistic morphology · usage data · split-100 task

The first author's work on this project was funded in part by the German Research Council (Deutsche Forschungsgemeinschaft, DFG, personal grant SCHA1916/1-1

Roland Schäfer
Deutsche und niederländische Philologie
Freie Universität Berlin
E-mail: roland.schaefer@fu-berlin.de

Elizabeth Pankratz
Deutsche Sprache und Linguistik
Humboldt Universität zu Berlin
E-mail: pankrate@hu-berlin.de

1 Linking elements in probabilistic morphology

In this paper, we examine so-called *linking elements* in German nominal compounds. Linking elements are optional segmental material inserted between the first and second nominal constituents of a compound, such as *-er* in *Liedertext* 'lyrics', literally 'song text'. Linking elements are not obligatory, and many compounds are acceptable with or without them. For instance, the variant *Liedtext* with zero linkage also exists, though speakers may have individual preferences for compounds with or without linking elements (as we will discuss further below). The distribution of the relatively large number of linking elements has previously been described in terms of constraints (some of them soft, others firmly categorical) from different angles (e. g., Fuhrhop 1996; Wegener 2003; Schlücker 2012; Nübling and Szczepaniak 2013; Fuhrhop and Kürschner 2015). Conceptually complete models of linking element selection have been proposed in rule-based systems (possibly with a similarity-based component; see Dressler et al 2001) and entirely exemplar-based models (Krott et al 2007). See also Section 2.

It is often assumed that linking elements have neither a grammatical function nor a semantic interpretation. In this paper, we question this perspective and look more closely at the possibility that certain linking elements could have a plural interpretation. Interestingly, except for several very rare linking elements, a default zero linkage, and the linking element *-(e)s*, all linking elements are formally identical to the first constituent's plural marker. Additionally, many first constituents alternate between such a pluralic linkage and another non-pluralic linkage (typically zero or *-(e)s*). This raises the question of whether speaker-hearers or writer-readers (or both) may associate a plural meaning with such pluralic linking elements. So far, researchers have been skeptical toward or even dismissive of this possibility, and in a perception experiment, Koester et al (2004) could not find evidence that a plural interpretation of linking elements is triggered in spoken German. However, for (written) Dutch linking elements, such effects have been demonstrated repeatedly (Schreuder et al 1998; Banga et al 2012, 2013a,b). In this paper, we present systematic research in the form of corpus data and the results of a split-100 experiment which shows that German linking elements are indeed used as cues for the plural meaning of the first constituent in written German compounds.

Findings like this have, in our view, an impact on morphological theory in general for two closely related reasons. First, any morphological framework or theory must be flexible enough to be able to account for empirically proven phenomena – in our case, the systematic occurrence of inflected forms within products of word-formation such as compounds. Classical generative frameworks (e. g., Siegel 1979; Mohanan 1986; Anderson 1992; Pinker 1999) tended to implement strong universal tendencies as hard constraints into the architecture of the framework, which, in our view, does not allow the required flexibility. One example is Siegel's strict layering hypothesis, which states that inflection applies derivationally after word formation and that consequently, markers of inflectional categories can only be positioned at the edges of words (as we will discuss further in Section 2.2.2). Such approaches have been criticised for many decades, since, as Haspelmath (2010, 391) puts it, "most empirical universals are tendencies". Pollard (1996) astutely criticises the absurdity of needlessly restrictive generative frameworks in syntax, and Haspelmath (2010) goes even further by suggesting that frameworks should abandoned altogether, demonstrating with many examples how frameworks sacrifice explanatory adequacy for hard-wired allegedly universal restrictions. With respect to inflection inside German compounds, we provide a framework-free but differentiated image to be summarised in Section 6.

Second, findings such as ours lend support to a probabilistic view of morphology and grammar in general. The amount of evidence for the inherent gradedness of grammar has been growing for decades. Hay and Baayen (2005) summarised an impressive number of studies about this topic as it concerns morphology, and Bresnan (2007) and subsequent work have radically changed the way empirical research is conducted in syntax. Even the relevance for linking elements has not gone uncommented; Arndt-Lappe et al (2016, 105) include the selection of linking elements in their list of "semi-systematic and gradient properties" of compounds. These properties have been actively researched in recent years, and the authors stress the important role played by empirical work in this area of investigation. The present study contributes to this body of work by showing that there is a – by no means categorical – tendency for linkages which take the form of a plural to be interpreted as plurals. Arndt-Lappe et al (2016, 107) address precisely this issue when they state that "although there is not and probably never has been a one-to-one correspondence between the form and meaning of compounds, the form does provide a wide variety of information to which humans have access in reaching an interpretation". As we will argue in Section 6, it would even be surprising if writer-readers did not pick up on the possibility of using plural forms to denote plural where it makes perfect sense, at least in the absence of strong inhibitory factors.

The paper is structured as follows. In Section 2, we review the form and distribution of linking elements in German, and we discuss positions about their potential plural interpretation from the literature. This includes a discussion of a long tradition of research on Dutch linking elements and their plural interpretability. The section closes with an outline of our hypotheses concerning when there might be a preference for pluralic linkages in compounds, for instance when the internal conceptual structure of the compound enforces a plural interpretation of the first constituent. Section 3 describes a database we created based on the DECOW16A web corpus to investigate these hypotheses. The database shows the frequencies with which a large number of nouns occurring as first constituents in compounds take pluralic and non-pluralic linkages. It also quantifies the productivity of each of these nouns with these linkages. In Section 4, we use the database to select a large number of first constituents which exhibit linking element alternation (in other words, first constituents which occur reliably with both pluralic and non-pluralic linkages) and show, using a manually annotated corpus sample of approximately $10,000$ compounds containing these first constituents, that pluralic linkages are indeed cues for semantic plurality. In Section 5, we report an experiment in the split-100 paradigm which corroborates our findings from the corpus study. Finally, in Section 6, we discuss the findings in the larger context of probabilistic morphology.

2 Linking elements in German

2.1 The form and distribution of linking elements in German

2.1.1 German linkages

In this paper, we exclusively deal with determinative (endocentric) nominal compounds, i. e., compounds formed through the concatenation of two constituent nouns (possibly with intervening linking elements, see below) where the second noun (N2) is the head modified by the first noun (N1). While compounding is a recursive process in principle, we ignore complications involved with more complex structures and look only at binary compounds. We refer to these structures as *N1+N2 compounds*.

In the majority of such compounds in German, N1 and N2 are simply concatenated, and N1 appears in its base (uninflected) form, which is identical to at least the nominative singular form. A linking element (LE), as mentioned above, is optional segmental material inserted between N1 and N2. LEs in compounds are not exclusive to German; see Schreuder et al (1998); Banga et al (2013a,b) on Dutch LEs.[1] However, German speakers use an unusually high number of different LEs, and they also make other types of changes to the base form of N1 (in order to refer to linking elements together with other formal changes to N1, we will use the umbrella term "linkage").[2] The list of segmental LEs is widely accepted to be: *e*, *er*, *s*, *es*, *n*, *en*, *ns*, and *ens* (see Neef 2015, 31, Krott et al 2007). *e* and *er* can occur together with an optional umlaut on N1, and some compounds only have umlaut on N1 and no segmental LE at all. Also, certain LEs can replace N1-final segments. Lastly, N1s ending in schwa (graphemically <*e*>) can drop this schwa in compounds.

In our analyses, we will not mark zero linkages in the simply concatenated compounds at all. *-X* is the notation we will use for a LE *X* attaching to N1, and *=X* denotes a LE attaching to N1 with additional umlaut. An umlaut-only linkage is indicated by = alone. The special notation **e* is used for a linkage where the final schwa of N1 is deleted. Furthermore, + separates N1 (possibly with LE) and N2 in our analyses. Graphemically, however, German compounds are usually spelled as one word; the segmentation is intended purely as a reading aid.[3] Table 1 illustrates the different types of linkages in German.[4]

Whether *-en* and *-n* and *-ens* and *-ns* should be described as allomorphic variants (i. e., *-(e)n* and *-(e)ns*) is debated. See Sections 3.2 and 3.6 of Nübling and Szczepaniak (2013) for an argument in favour of this view and Neef (2015, 33–36) for an argument against it. This issue does not affect our study, and we separate the two potential allo-forms in order

[1] See also Krott et al (2007, 27) for a typological overview of some systems of compound infixoids.

[2] Linkages in German have developed diachronically from inflectional forms, see Nübling and Szczepaniak (2013); Szczepaniak (2016).

[3] Some compounds, for example those involving names or loan words, are sometimes written with a hyphen, such as *Adenauer-Zeit* 'Adenauer period'. We exclude them from our study because they virtually never occur with LEs and have specific compound-internal semantics. There are also occasional spellings of compounds as two words such as *Blumen Welt* 'world of flowers' and with in-word capital letters such as *BlumenWelt*, which are much debated from a normative perspective (see also Scherer 2012). These spellings are assumed to be specific to certain genres (such as adverts) and are quite rare overall. Therefore, we do not include them in our study.

[4] Different types of linkages are also found in other sorts of German compounds, for example N+A compounds such as *herzensgut* (*Herz-ens+gut*) 'kindhearted', *lebensfreundlich* (*Leben-s+freundlich*) 'life-sustaining/livable', or *hundelieb* (*Hund-e+lieb*) 'dog-loving'. While their distribution, function, and interpretation in such compounds might be related to that in N1+N2 compounds, our studies reported in Sections 4 and 5 make use of diagnostic features exclusive to N1+N2 compounds, and we therefore excluded all other types of compounds.

Linkage	Pl	example	literal gloss	translation
∅		*Haus-tür*	house door	'front door'
-s		*Anfang-s+zeit*	beginning time	'initial period'
-n	✓	*Katze-n+pfote*	cat paw	'cat's paw'
-en	✓	*Frau-en+stimme*	woman voice	'female voice'
*e		*Kirsch+kuchen*	cherry cake	'cherry cake'
-e	✓	*Geschenk-e+laden*	gift store	'gift store'
=e	✓	*Händ=e+druck*	hand press	'handshake'
=	✓	*Mütter=+zentrum*	mother centre	'centre for mothers'
-er	✓	*Kind-er+buch*	child book	'children's book'
=er	✓	*Büch=er+regal*	book shelf	'bookshelf'
-ns		*Name-ns+schutz*	name protection	'name protection'
-ens		*Herz-ens+angelegenheit*	heart matter	'affair of the heart'

Table 1 Overview of the different linkages in German nominal compounds; = is used to denote LEs which trigger umlaut on N1, **e* denotes linkages where the final schwa of N1 is deleted; the column labelled *Pl* indicates whether the linkage is (in most cases) formally identical to the plural of N1 (see below)

to maximise the informativity of our results. Since we do not look at linkages with *-s* and *-es* in detail, however, we have conflated these two into *-(e)s*. The different linkages have quite different frequencies, with the zero and *-(e)s* linkages being by far the most frequent ones. For example, Gallmann (1998, 177) reports that 70% of all N1+N2 compounds have zero linkage (not specifying whether he refers to type or token frequency). Krott et al (2007, 29) report that the compound types with zero linkage make up 65% of all compounds in the CELEX database. See Section 3.2 for a detailed breakdown of the type and token frequencies in the 21 billion token DECOW16A corpus.

2.1.2 Conditions on linkage selection in German

The choice of linkage in any given compound is not fully predictable, but the grammatical gender of N1, its declension class (dependent largely on gender), its phonotactics, and possibly plural semantics all limit the options significantly (see Fuhrhop 1996; Nübling and Szczepaniak 2013). All these morpho-phonological and lexical factors which partially determine the choice of linkage are based on N1, so the distribution of linkages in general is unanimously treated as suffixation of N1.

Several (soft) descriptive generalisations concerning linkage distribution can be made. Derived nouns ending in suffixes like *-ung*, *-heit*, *-tum*, etc. have a strong tendency to occur with an *-s* linkage, as in *Heizung-s+wartung* 'heating maintenance'. Simplex masculine and neuter nouns are those which often take the *-(e)s* LE, in which case the form is identical to the genitive singular, for example *Boot-s+fahrt* 'boat trip' (with neuter *Boot*) or *Tag-es+form* 'form of the day' (with masculine *Tag*). Also, N1s ending in full (non-reduced or tense) vowels such as *Oma* 'grandma' or *Auto* 'car' can never take the *-s* LE (although *-s* is their only inflectional suffix when they are used as independent nouns) and tend to occur with a zero linkage as in *Auto+wäsche* 'car wash'. See Wegener (2003) and Fuhrhop and Kürschner (2015) for further discussion of the *-s* linkages and Fehringer (2009) for an investigation of an emerging *-s* PL in northern German dialects. Furthermore, feminine nouns ending in schwa virtually always occur with the linking element *-n* (Libben et al

2002, 32), which is also their plural morpheme. So-called weak masculine nouns and so-called mixed masculine and neuter nouns, which have an *-(e)n* plural, often occur with an *-(e)n* LE, e. g., *Linguist-en+witz* 'linguist's joke/joke about linguists' and *Schwede-n+humor* 'Swede humor'.[5] Finally, *-ens* and *-ns* are idiosyncratic and rare, used only with a handful of N1s.

More systematically, Dressler et al (2001) showed in a cloze test that the choices German native speakers make with respect to linkages in novel compounds are partially predictable from rules, but that an analogical component (similarity to existing compounds with the same first constituent) is required. The approach in Krott et al (2007) shows how a model with very high explanatory power can be constructed completely without rules and exceptions, based only on analogy. Using an implemented exemplar-based model, experimental validation, and post-hoc analysis, they find that the choice of linkages in novel compounds can be predicted well by considering exemplar families of compounds that have the same N1 together with features of the first constituent such as rime, gender, and inflectional class (Krott et al 2007, 47).

In sum, the zero linkage can be considered a sort of default, since it is the most frequent form of linkage in N1+N2 compounds and no compound is phonologically unacceptable with a zero linkage. When a LE does appear, which form it takes is affected primarily by N1. N1s have certain preferences for linkage selection, and they also often disprefer particular linkages (partially depending on gender, stem-final segments, and the derivational status of N1). There is solid evidence that exemplar effects can account for most selection tendencies. It remains a noteworthy fact, however, that except for the zero and *-(e)s* linkages, the linkages which N1s can take are almost always identical to their plural markers.

2.1.3 Alternations between linkages

Despite the existence of more or less firm constraints on the selection of linkages, which linkage will be chosen in a given compound is by no means fully pre-determined. Dating back to Augst (1975), alternations between one or more different linkages with the same N1 have been described. Alternations with more than two alternatives are rare, however. Augst (1975, 134–135) reports that among the 4,025 N1s he examined, 390 occurred with two linkages, 31 with three, and only eight with four. Notice that the proportion of N1s occurring with two different linkages is considerable at 9.7%. We propose that corpora containing more liberal usage of language which is not strongly bound by norms combined with contemporary technologies of large-scale automatic analysis will reveal many more N1s occurring in at least a two-way alternation between a zero linkage and a pluralic linkage (see Section 3 and Section 4).

Some researchers are skeptical towards the idea of productive alternation. Roughly forty years after Augst's study, Neef and Borgwaldt (2012, 31) and Neef (2015, 46) suggest that for each N1, there is one uniquely determined linkage, and they treat all cases of alternation as lexicalised "exceptions". In Neef and Borgwaldt (2012, 42), the authors carry out what they call a "corpus study" based on the normative spelling dictionary *Duden* (Dudenredaktion 2006). From the 19 compounds they found which contain *Ohr* 'ear' as N1 with a pluralic linkage (*Ohr-en*) and the 15 compounds with *Ohr* and a non-pluralic (zero) link, they conclude that "the method does not result in a clear picture about the productive form of the

[5] For weak nouns, the *-(e)n* LE is also identical to all non-nominative-singular forms, because they follow an otherwise unusual paradigm where the nominative singular is unmarked and all other forms are marked identically with *-(e)n* (see Köpcke 1995; Schäfer 2016).

first constituent" and that there might simply be a "a developing allomorphy of the first constituent for this particular lexeme" to explain the similar compound type frequencies of both linkages.[6] In light of contemporary methods in corpus linguistics, such an argumentation can be dismissed on purely methodological grounds. We describe a principled approach to assessing productive linkage alternations for N1s in Section 3.

2.2 Linkages as plural markers

2.2.1 Plural interpretation of pluralic linkages in German and Dutch

Our study focusses on N1s which alternate between a pluralic and a non-pluralic linkage, and it is our goal to show that these linkages actually have a plural interpretation and to clarify when this appears. If they do have a plural interpretation, then at least in some compounds, the linkage is a case of plural inflection inside a compound. We now review some of the research on a potential plural interpretation of pluralic linkages in German and Dutch.

An extreme position with respect to any functional or semantic interpretation of linkages is espoused by Neef and Borgwaldt (2012) and Neef (2015). In these papers, the authors focus on the (undisputed) fact that none of the functions attributed to different linkages (see Section 2.1), including a potential function as a plural marker, apply strictly across the board. The plural marking function is not strict insofar as a pluralic linkage is not required for N1 to be interpreted as a semantic plural, and a pluralic linkage also does not reliably exclude a singular interpretation. While this does in no way exclude the possibility that a pluralic linkage is a soft cue for plurality, the authors seem to suggest that function/meaning and form have to stand in a one-to-one relationship, except for a small number of exceptions which have been lexicalised (e. g., Neef and Borgwaldt 2012, 42). They dismiss polyfunctional (and probably multifactorial) explanations altogether, because no stringent system of functions has been proposed which explains exactly under which conditions which linkage is chosen (e. g., Neef and Borgwaldt 2012, 27–29). Such a view (similar to the dual-route approach as discussed in the context of LEs in Krott et al 2007) has become less and less tenable in the face of recent research on the nature of the form–meaning relationship in morphology, which paints a clearly probabilistic picture where more often than not, similarity relations play a major role (see Arndt-Lappe et al 2016, 107). The present study adds evidence that axiomatic and/or Aristotelian approaches are inadequate with regard to the interpretation of German linkages.

More substantially, it was shown in Koester et al (2004) that in the perception of spoken German compounds, hearers do not use pluralic linkages as cues to plural semantics. However, much like Schreuder et al (1998) argue for effects of plural semantics in written Dutch compounds (see below), the picture might be different for written German.

Concerning the possibility of inflection within compound boundaries, Schlücker (2012, 9) states that there can be no inflection on non-heads in compounds, because these non-heads either do not inflect at all or the suffixes with which they occur (i. e., LEs) are non-inflectional by definition. In their survey of linkages in Germanic languages, Fuhrhop and Kürschner (2015, 577) state more liberally that "the expression of number by linking elements generally seems possible", while case marking, another instance of inflection, is mostly considered to be unavailable within compounds.

[6] "Damit liefert die vorgeschlagene Methodik kein klares Bild zur produktiven Vordergliedsstammform. Denkbar ist allerdings, dass sich für dieses spezielle Lexem eine Allomorphie auf der Ebene der Vordergliedsstammform herausgebildet [. . .]" (Neef and Borgwaldt 2012, 31)

Gallmann (1998, 178–180) distinguishes between internally licensed and externally licensed inflectional features. He adopts a specific formulation within the Government and Binding framework, but we focus on the gist of his generalisation, which can be expressed well outside of that specific framework. Externally licensed inflectional features are those which are determined by the syntagmatic context in which a noun occurs, such as case assignment by verbs. Internally licensed features, on the other hand, are those which are assigned not by context, but by categorial membership (such as grammatical gender) or interpretation (such as free plurals).[7] Gallmann maintains that externally licensed inflectional features – prominently, case – cannot be assigned to non-heads in compounds (N1s), but he does not explicitly exclude internally licensed features like plural marking on N1s.

These examples, though far from all comments made on this topic, illustrate the divide among linguists and grammarians of German. Considering this divide, astonishingly little empirical research has been published on the question. However, for Dutch, there exists a decades-long tradition of substantive experimental research into plural interpretations of pluralic linkages. Schreuder et al (1998) and Banga et al (2012) demonstrate how a change in the official Dutch orthography, which assimilated the LE to the plural marker graphemically, fostered the plural interpretation of non-heads of compounds in written Dutch. Also, Banga et al (2013a) show that a plural interpretation of N1s in novel Dutch compounds is positively linked to the occurrence of the optional LE *-en*, which is homophonous with the plural marker. First, they showed that subjects prefer to use *-en* with N1 in contexts where a plural meaning of N1 was made clear. Second, they demonstrate that the preference for the pluralic linkage is also activated when the context creates a plural meaning for N1 but contains only a singular form of N1, which is evidence that the connection between the LE and plurality is not just based on a formal recency effect. They confirm that plural interpretation indeed creates a preference for using *-en*, but that form-based repetition effects strengthen the meaning-based effect (Banga et al 2013a, 45). Finally, they also test German L2 learners of Dutch and find that for them, the effect is weaker than for native speakers of Dutch. They offer two possible interpretations for this (Banga et al 2013a, 45–47). Citing Libben et al (2002), who show that LEs in German come with a processing overhead, they speculate that German speakers might have a tendency to avoid LEs. Also, they propose that since German plurals are not always marked (there is a zero plural in German, just as there is a zero linkage), German speakers might associate linkages less strongly with plurality. Finally, they argue that plural markers in German are sometimes formally identical to case markers, which could also weaken the connection German speakers establish between plurality and linkages. We return to these ideas in Section 6 after having presented our own results.

In Banga et al (2013b), the authors compare the interpretation of Dutch compounds by native speakers of Dutch to the interpretation of conceptually identical English compounds by English native speakers. They compare plurality ratings for conceptually plural N1s, i. e., those where a plural meaning of N1 is a de-facto necessity, to plurality ratings for conceptual non-plurals. Their examples include *bananenschil* ‘banana peel’ for a conceptual non-plural and *aardbeienjam* ‘strawberry jam’ for a conceptual plural. In essence, they show that Dutch speakers produce higher plurality ratings for compounds with *-en* and lower plurality ratings for compounds without *-en* for both groups of compounds (conceptually plural and non-plural) compared to English native speakers for the English compounds identical in meaning and structure. That is, the LE provides Dutch speakers with an additional cue for plurality, independent of the cues coming from the conceptual structure of the compound.

[7] An example of a non-free plural according to Gallmann (1998, 179) would be the plural on *students* in *Dale and Daryl are students*. See his parallel German example (4b).

In a second experiment, they show that Dutch subjects as L2 speakers of English react to English compounds by and large like English speakers (for a report of some complications, see Banga et al 2013b, 211). This corroborates findings that the actual presence of the LE is a key cue for plural interpretation for Dutch speakers.

Finally, it should be mentioned that in some terminologies, LEs would not be called LEs when they are actual plural markers (i. e., when plural form of the LE and plural meaning of N1 occur together). We follow the liberal choice of words by Banga et al (2013b), who also call the respective affixes "linking elements" whether or not they mark plural meaning (see Banga et al 2013b, 196 for a discussion of stricter views).

2.2.2 Consequences of plural linkages for morphological theories

A crucial point in the discussion of inflection within compounds is whether frameworks and theories can deal with it appropriately, as we mentioned above. In theories ranging from the strict layering hypothesis by Siegel (1979) through the lexical Phonology of Mohanan (1986) and the a-morphous morphology by Anderson (1992) to the words-and-rules theory by Pinker (1999), it has been maintained that inflection applies after derivational word formation (including compounding), leading to inflectional affixes being positioned farther toward the edges of words than derivational affixes and to inflectional affixes not occurring between constituents of compounds. This was encoded in the respective theories as hard and putatively universal constraints. Such approaches were disputed, however, from very early on, for example in Bochner (1984), where the author shows how inflection can occur within derivation (see Kirchner and Nicoladis 2009, 2–3 for more on this debate). Most pertinent in the context of our study is the argument in Banga et al (2013a, 47–48) that such restrictive theories fail to explain how the Dutch LE *-en* can have a plural interpretation (thus being an inflectional suffix) and occur between constituents of compounds (see their results summarised above). The same would be true for German compounds if pluralic linkages turn out to be plural markers systematically.

However, as Kirchner and Nicoladis (2009, 5) correctly point out, strong tendencies (falsely interpreted as universal constraints in older generative morphology) for inflection to appear only at the edges of words should still be explained. Given this goal, we consider it highly instrumental to describe exactly where and when inflection does *not* appear at the edges of words, and this paper is about such a case (see Section 6). As we have argued in Section 1, frameworks should not impose unnecessary restrictions which stand in the way of describing phenomena as they actually occur. Furthermore, the inherent probabilistic nature of linguistic phenomena – most obvious in situations where more than one option is available in a so-called *alternation* – means that investigating preferences might be more useful than searching for strict rules and lexicalised exceptions.

2.3 Outline of our study

Our main hypothesis under investigation is whether a pluralic linkage in a N1+N2 compound is systematically interpreted as a semantic plural by native speakers of German. We use the term "systematic" in the sense of probabilistic grammar, where preferences (under specific given conditions) are not assumed to be binary but probabilistic, weighted, and better described as numerical rather than discrete. As sources of data, we will use a large corpus and an experimental setup. Since annotating corpus exemplars reliably for whether the N1 in a compound has a plural meaning is quite difficult in the general case, we isolate two

specific and relatively easy-to-detect configurations in which a pluralic linkage might signal plurality, where the second configuration (internal plural) is the one we consider to be the crucial one in showing that pluralic linkages have plural meaning:

1. A plural on the entire compound (formally on the head constituent) might trigger the use of a pluralic linkage. We call this the *external plural effect*.
2. Certain semantic classes of N2s standing in an appropriate semantic relation with N1 might force N1 to have a plural interpretation and therefore lead to a preference for using the pluralic linkage. We call this the *internal plural effect*.

The external plural effect on linkage selection might have two different motivations. First, there are cases where the referents of N1 necessarily form a set or sum entity with more than one member when the compound as a whole is a plural. This might lead to a preference of pluralic linkages. For example, we might see a preference for a compound *Hund+herz* 'dog's heart' with the non-pluralic linkage (zero in this case) in the singular but *Hund-e+herzen* 'dogs' hearts' with the pluralic linkage (*-e* in this case) in the plural. It is not necessary to distinguish between different non-pluralic linkages (mainly zero and *-(e)s*). This effect depends to a large extent on the semantics of both constituents and the compound. In the case of *Hundherz*, the effect is clear because each dog has exactly one heart, so multiple dogs entail multiple hearts. With other compounds, such as *Brot+mahlzeit* or *Brot-e+mahlzeit* 'bread meal', the picture is blurrier because a single loaf or piece of bread can make several meals, and more than one piece or loaf of bread can be consumed in one meal. Making matters worse, it could be the generic/mass noun meaning of *Brot* which is addressed, in which case plurality does not even make sense. Since we could not find a way to solve these problems reliably with manual annotation, we treat the external plural effect on a strictly formal level.

The second potential motivation for the external plural effect is a purely formal one; there might simply be plural agreement within the compound. Banga et al (2013a) and Banga et al (2013b) have found that there are effects related to the mere presence of a formal plural of N1 in the context of the compound. However, the plural on the whole compound might trigger a preference for a pluralic linkage on N1, even if the semantic motivation for the external plural effect does not apply. Thus, even if we find evidence for an external plural effect, we could not be sure that it is an effect related to plural meaning.

The internal plural effect, on the other hand, is only related to the lexical meanings of the compound's constituents, and it is similar to the conceptual plurals in Banga et al (2013b), as discussed above. Prominently, N2s that provoke this effect might have a collective meaning. In this case, regardless of the grammatical number of the whole compound, there are necessarily several referents of N1 involved conceptually. Examples include true collectives like *Kindergruppe* 'group of children', metaphorical collectives as in *Zitateregen* 'rain of quotations', reciprocals such as *Räderwechsel* 'swapping of tyres', or relational N2s as in *Lochdistanz* 'distance between (the) holes'.

The external plural condition is weaker both conceptually and in terms of operationalisation, so if pluralic linkages are indeed interpretable as plural markers, we expect to find an effect especially for the internal plural condition. Interestingly, if we find evidence only for the internal plural effect but not the external plural effect, this would fit within a probabilistic version of previous findings that plurality in compounds is purely conceptual or inherent to N1 and does not depend on the grammatical context of the compound (see discussion above, for example Gallmann 1998).

In Section 4, we study these two specific hypotheses using corpus data. Then, in Section 5, we examine them in an experimental paradigm (split-100 ratings). Before we turn to these studies, Section 3 describes the exploratory work that allowed us to make an informed selection of items for the studies. It also shows that there is a significant number of N1s alternating between pluralic and non-pluralic linkage, contrary to statements to the contrary in the literature (most prominently Neef and Borgwaldt 2012; Neef 2015).

3 Data

3.1 Corpus choice

We chose the web-crawled DECOW16A corpus (Schäfer and Bildhauer 2012, Schäfer and Bildhauer in prep.) for all our corpus work.[8] This corpus was the obvious choice for several reasons. First of all, it is available in an on-line query interface but also for scripted access and (in sentence-wise shuffled form) for download.[9] The large-scale productivity assessment to be reported in Section 3.2 would not have been possible without scripted access. More importantly, using large amounts of recently produced data, including data not written under strong normative pressure (such as text from forums and other community websites) is in our view ideal for research on productive processes from a synchronic perspective (be it descriptive, geared towards competence grammar, or cognitively oriented). The only other available very large corpus containing recent German would be the *Deutsches Referenzkorpus* (DeReKo) by the *Institut für Deutsche Sprache* (Kupietz et al 2010), but (at least currently) it is not available for scripted access, and it mostly contains newspaper text. Finally, the COW corpora are based on an improved methodology also used to build the WaCky corpora (Baroni et al 2009), and there are other similar web-derived corpora also actively used by many linguists, such as the SketchEngine corpora (Kilgarriff et al 2014). Web corpora can thus be regarded as an established source of data on par with traditionally compiled corpora.

3.2 Database and productivity assessment

For the studies reported in Section 4 and 5, we had to choose a set of first constituents (N1s) to be examined more thoroughly using corpus data and a set of compounds to be used as stimuli in the split-100 experiment. The total number of candidates for first constituents and compounds is quite high (from thousands to tens of thousands), and in order to make an informed selection, large-scale data about the type frequencies, token frequencies, and the productivity of first constituents were required. We used automatic approaches to create a set of databases, and manual clean-up and selection steps in-between automatic steps were used to ensure that the results were reliable. The database tells us how strongly first constituents tend to appear with PLs and with NPLs, including an assessment of their productivity with these linkages. This section describes the creation of the database.

We began by extracting a very large database of all nominal compounds (not just N1+N2 compounds but any compound with a nominal head) from the DECOW16A web corpus (Schäfer and Bildhauer 2012, Schäfer and Bildhauer in prep.).[10] In the corpus, nominal compounds come with full structural analyses created automatically using the SMOR finite-state morphological analyser (Schmid et al 2004) and extensive pre- and post-processing im-

[8] See `http://corporafromtheweb.org/` for project information and `https://www.webcorpora.org/` for access to the corpora.

[9] It is also available free of charge to anyone working in academia. The same is also true for the English ENCOW16A (16.5 billion tokens), the French FRCOW16A (10.8 billion tokens). the Spanish ESCOW16A (7.1 billion tokens), as well as the older Swedish SVCOW14A (8.4 billion tokens) and Dutch NLCOW14A (6.7 billion tokens).

[10] In agreement with the creators of the DECOW16A corpus, our database, which contains comprehensive aggregated information about the 22,380,133 compound types accounting for 478,342,305 tokens in the corpus, will be made publicly available.

plemented by the COW creators (Schäfer and Bildhauer in prep.).[11] While we noticed that the automatic analyser (like all automatic annotation tools) makes some errors and sometimes fails in disambiguating ambiguous compounds, it also became clear that the quality of analysis is more than sufficient for the kind of large-scale pre-analysis we performed. More importantly, at each step of the analysis, we made sure through manual checks that the data which we actually used for the studies were clean.[12]

	Compound *F*	%	Compound *f*	%	N1 *F*	%
∅	3,409,883	60.252	194,187,343	61.379	18,515	41.769
-s	1,340,565	23.687	79,131,193	25.012	20,274	45.737
-n	587,365	10.379	25,175,402	7.957	2,343	5.286
-en	170,906	3.020	5,293,805	1.673	1,858	4.192
*e	43,876	0.775	6,744,777	2.132	88	0.199
-e	39,978	0.706	1,560,226	0.493	844	1.904
-er	21,398	0.378	2,109,421	0.667	30	0.068
=er	20,679	0.365	893,547	0.282	50	0.113
=e	12,679	0.224	578,171	0.183	279	0.629
-ns	7,205	0.127	546,419	0.173	16	0.036
=	3,046	0.054	63,591	0.020	28	0.063
-ens	1,807	0.032	90,428	0.029	2	0.005
Σ	5,659,387	100.000	316,374,323	100.000	44,327	100.000

Table 2 Type and token frequencies of all linkages in N1+N2 compounds in DECOW16A

First of all, as argued for in Section 2, we restrict our study to N1+N2 compounds. We therefore extracted all N1+N2 compounds and all first constituents (N1s) from the large compound database. For each of the pluralic linkages (PL), we generated exhaustive lists of the N1s with which it occurs which have a plural identical to the PL (to exclude the few cases where a potentially PL attaches to a noun of a different declension class), which are count nouns (because mass nouns undergo a significant change in meaning when pluralised) and which are not weak masculine nouns (because these do not differentiate between singular non-nominative forms and plural forms). The lists of N1 candidates were extracted automatically, but they were checked for the aforementioned criteria and sporadic results of incorrect automatic analysis manually. Both the first author of this paper and a student assistant checked the entire list in order to increase the accuracy of the manual clean up process.[13] When an N1 candidate looked suspicious, we inspected the compound database

[11] The analyses are formatted as `Zeit_Punkt` for *Zeitpunkt* 'point in time/moment' (literally 'time point') or `Wort_+=er_Buch` for *Wörterbuch* 'dictionary' (literally 'word book'). Underscores separate the constituents of the compound, the nominal elements are given in their base form, and LEs begin with + followed by a = if the LE triggers stem umlaut on the first constituent.

[12] We made an exception for the zero linkage because the number of N1s was much too high for full manual inspection. We therefore implemented an automatic approach where a N1 had to fulfil one of the following criteria to be included: (i) It was a word known to the aspell spell checker (in the mixed capitalisation typical of German nouns) and not in the list of proper nouns extracted from the DECOW corpus published by the COW project. (ii) It occurred at least 50 times in DECOW16A in mixed capitalisation spelling and was not in the list of proper nouns. These criteria led to 4,473 of 22,988 nouns being excluded from the list of zero linkage N1 candidates.

[13] Every single decision is documented in the data package for this paper.

to check whether the actual compounds containing the candidate were misanalysed or otherwise noisy.

In total, we found 4,393 different N1s which are used with a PL. Then, for each N1 we counted the numbers of compound types, the number of compound tokens, and the number of compound hapax legomena (i. e., those compounds containing the N1 which occur only once in the corpus) containing it. Since we are interested in the alternation for each N1 of PLs and non-pluralic linkages (NPLS; these can manifest themselves as no LE at all, as deletion, or as a non-pluralic LE such as *-s*), we also extracted the same counts for the N1s in compounds with a NPL. An overview of the type frequencies (F) and token frequencies (f) of the different linkages in DECOW16A is given in Table 2.[14] The numbers are approximately in line with previous reports based on much smaller corpora such as Wellmann et al (1974), Kürschner (2005), or Krott et al (2007) (see also Schlücker 2012, 9). As explained in Section 2, the classes of words with the plural morphemes *=er*, *-er*, and *=* are rather small with 50, 30 and 28 different N1 types in our database, respectively. While *=e* has an intermediate type frequency of 279, *-n* (2,013 types), *-en* (1,149 types), and *-e* (844 types) are highly type-frequent.

While raw type and token frequencies are informative to some extent, a comparison of their productivity with PL and NPL is equally important for making decisions regarding the corpus sample and selection of stimuli. Therefore, we calculated measures of productivity for each N1 with PL and NPL. This measure was supposed to capture the probability that the given N1 would form a new compound with PL and NPL. The *potential productivity* is the measure of choice for this as it "gauges the extent to which the market for a category is saturated" (Baayen 2009, 902, see also 906–907). The potential productivity is appropriate for our purposes because some compounds might be lexicalised either with PL or with NPL, and it would not make much sense to try to examine an alternation with N1s which occur mostly in lexicalised compounds. As shown in (1), the potential productivity P^p is simply the number of the hapax legomena (f_1) of compounds with $\mathrm{N1}_{le}$ divided by the token frequency (f) of compounds with $\mathrm{N1}_{le}$, where *le* stands for either PL or NPL.

$$P^p(\mathrm{N1}_{le}) = \frac{f_1(\mathrm{N1}_{le})}{f(\mathrm{N1}_{le})} \tag{1}$$

The interpretation of P^p is intuitive, as it is 0 when there are no hapax legomena in the corpus ($f_1(\mathrm{N1})_{le} = 0$; no productivity whatsoever), and it is 1 if all tokens are hapax legomena ($f_1(\mathrm{N1})_{le} = f(\mathrm{N1})_{le}$; maximal productivity). It can be regarded as a proportion, and its range is therefore $[0, 1]$.

[14] In these counts, mass nouns and weak nouns were included in order to provide a complete overview. Since they were removed for all further analysis reported below, the frequencies in the table are higher, especially for *-n* and *-en*, which are the hypothetical PLs of the weak nouns.

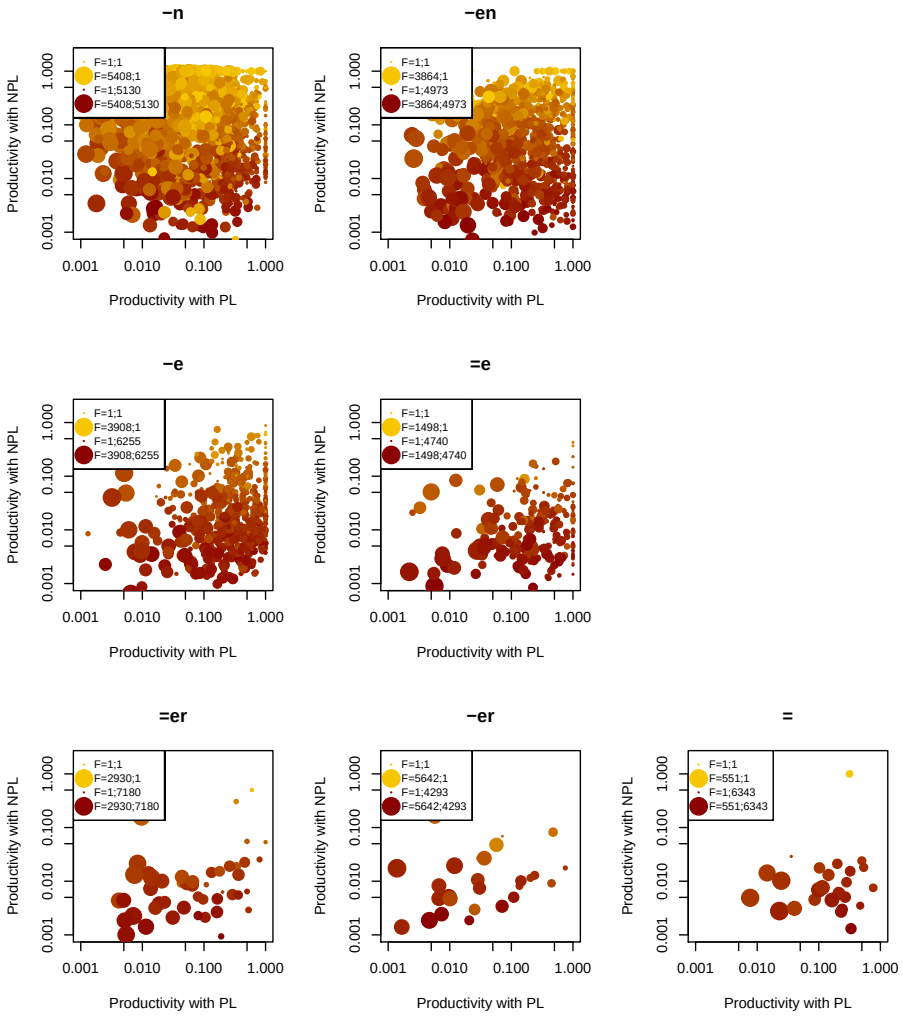

Fig. 1 Comparisons of the potential productivity of N1s, grouped by PLs; colours and sizes encode the type frequencies with PL and NPL; axes are on a logarithmic scale

Figure 1 shows the results of the productivity analyses for all chosen N1s.[15] In this plot, each dot represents one N1. The dot's position is determined by its P^p value with PL (x-

[15] It was pointed out that the analysis of productivity is often tainted by the low quality of automatic annotation in Evert and Lüdeling (2001) – a point reiterated in Baayen (2009, 907). Our analysis relies on the automatically generated annotations by the SMOR tool and COW tool chain. However, the list of N1 candidates was meticulously cleaned manually, as stated above. Also, we have found that the compound analyses by the SMOR tool are highly accurate when it does not have to guess the lemma of a compound's elements, and the N1s actually used are obviously known words, since we input them. Finally, the N1s which were chosen for the studies presented in Sections 4 and 5 were manually checked again. In the annotation of the 9,414 corpus exemplars, any systematic misanalysis would have shown. This actually was the case for the N1 *Beere* 'berry', where the concordance for NPL contained mostly erroneously analysed words and proper names, and *Beere* was subsequently excluded from further steps. We can thus be sure that the analyses which

axis) and with NPL (y-axis). Additionally, the larger a dot is, the higher is its type frequency with PL, and the darker it is, the higher is its type frequency with NPL. From the panels for *-n* and *-en*, it is apparent that N1s with all sorts of ratios of high and low productivity with PL and NPL exist. The tendency for dots to be smaller towards the right-hand side (high productivity with PL) and lighter towards to top (higher productivity with NPL) is explained by the fact that a lower overall type frequency makes it is easier to achieve a high productivity score. In the extreme case, a N1 has a type frequency of 1, and there is only one occurrence of it (necessarily a hapax legomenon), which results in a potential productivity score of 1. Since the *-n* and *-en* plurals are often used with rare loan words, there are many items with low type frequency and a high productivity score. Examples include *Ikonostase* (pl. *Ikonostasen*) 'iconostasis', which has a token frequency and a hapax count of 2 with PL (thus a potential productivity of 1) and *Testator* (pl. *Testatoren*) 'testator' with a token frequency and a hapax count of 1 with PL (thus also a potential productivity of 1).

For N1s with *-e* and *=e*, and even more so for those with *=er*, *-er*, and *=*, the productivity scores with PL are spread out between 0 and 1. However, there are virtually no N1s in these classes which show a particularly high productivity (much higher than 0.1) with NPL. The type frequencies are, however, still quite high for both PL and NPL, as reflected in the colour and the size of the dots. For *=er*, for example, the 25th and 75th percentiles of the token frequencies with PL are at 16 and 469 types, respectively. With NPL, they are at 482 and 2,190 types. In Sections 4 and 5, we use the data described here to make informed selections of items for further study and we also detail the studies performed with these items.

4 Corpus study

4.1 Queries

As explained in Section 2, we will determine whether two distinct factors increase the probability that N1 in an N1+N2 compound occurs with a PL, given that N1 alternates between a PL and a NPL. The first potential factor is a plural on the whole compound (formally on N2, which is the head). The second factor is whether a semantic relation between N1 and N2 holds which forces N1 to have plural semantics. Therefore, we manually annotated corpus exemplars containing N1+N2 compounds for whether they are plurals and whether a plural-enforcing semantic relation holds between N1 and N2. Before turning to this annotation process in Section 4.2, this section describes how we extracted and prepared a concordance for the manual annotation.

First, a selection of N1s was required which represents the population of N1s well with regard to their productivity scores with PL and NPL. A set of criteria was devised to make sure that that was the case. We required that the potential productivity score with PL and with NPL not be 0 or 1, which excluded rare words. Furthermore, only N1s which had a minimal type frequency of 50 as N1 with PL and NPL were used. In order to exclude nouns which occur with reasonable type frequency in compounds but infrequently as standalone nouns, only N1s with a minimal token frequency outside of compounds were included. To implement this restriction, we used the notion of the *frequency class* (or *frequency band*) of a word (see Perkuhn et al 2012, 80). The frequency class $c(w)$ of a word w increases with the word's token frequency $f(w)$. Calculation of the frequency class additionally relates the

we interpret relative to our theoretical hypotheses are accurate and not distorted by problems of automatic annotation.

token frequency of the word in question to the number of tokens of the most frequent word in the corpus (f^{max}) and accounts for the power law distribution of word frequencies.[16] It is given by (2), where $\lfloor\ \rfloor$ denotes the function that rounds down to the next integer.

$$c(w) = \left\lfloor 0.5 - log_2 \left(\frac{f(w)}{f^{max}} \right) \right\rfloor \tag{2}$$

The most frequent word in DECOW16A, according to the official frequency lists, is *und* 'and' with $f(und) = 258,507,195$. Relative to this f^{max} value, we only considered words with a frequency class up to and including 15, a class where, for example, *Lid* 'eyelid' ($f = 10,746$), *Seilschaft* 'rope team' ($f = 9,237$), and *Verlies* 'oubliette' ($f = 6,734$) are found.

From the set of N1s which fulfilled the given criteria, we sampled between five and ten nouns per PL. Sampling was a manual process, and it was ensured that the nouns were countable (i. e., no mass nouns), not collectives, and distributed approximately uniformly across the productivity spectrum as it was visualised in Figure 1. In Figure 2, each of the dots represents one N1 which fulfils the selection criteria. The triangles mark those which were chosen for the corpus study. It is immediately obvious that the sample represents the overall spectrum of productivity quite well.

The queries we used to retrieve the exemplars of compounds containing the fifty chosen N1s were made in DECOW16A using Python through the RStudio Server interface provided by the COW project.[17] For each N1, we made one query searching for compounds containing it with PL and one otherwise identical query searching for compounds containing it with NPL. Since at least some context is often needed to disambiguate the meaning of the compounds and their constituents, the whole sentence containing each compound as well as one sentence to the left and one sentence to the right were exported. Duplicate sentences were removed by setting the appropriate parameter in the COW project's Python wrappers for making queries. Additionally, we performed deduplication on the compounds inasmuch as we only allowed one instance of each compound word form to pass. Otherwise, highly frequent and typically lexicalised compounds would have accounted for the best part of the concordances. However, we did allow different word forms of the same compound in the concordances, in order to get both singular and plural forms. We extracted all exemplars matching the query, making random sub-samples for annotation.[18] In addition to the exemplar in its context, we extracted the corresponding document URL, the unique COW16 ID of that document, and the sentence ID. With the COW16 ID and the sentence ID, each exemplar can be located in the DECOW16A corpus for full reproducibility.

4.2 Annotation

From each of the query results of the fifty N1s described in Section 4.1, we annotated approximately one hundred cases with PL and approximately one hundred with NPL. Due to some minor clean-ups in the quality control process, the final sample size is not exactly $n = 9,600$ but $n = 9,414$.

[16] See Piantados (2014) for a recent overview, including the many problems with actual word frequency distributions.

[17] The Python scripts showing the exact queries will be published openly as part of the data package for this paper.

[18] The full query results are also included in the data package which accompanies this paper.

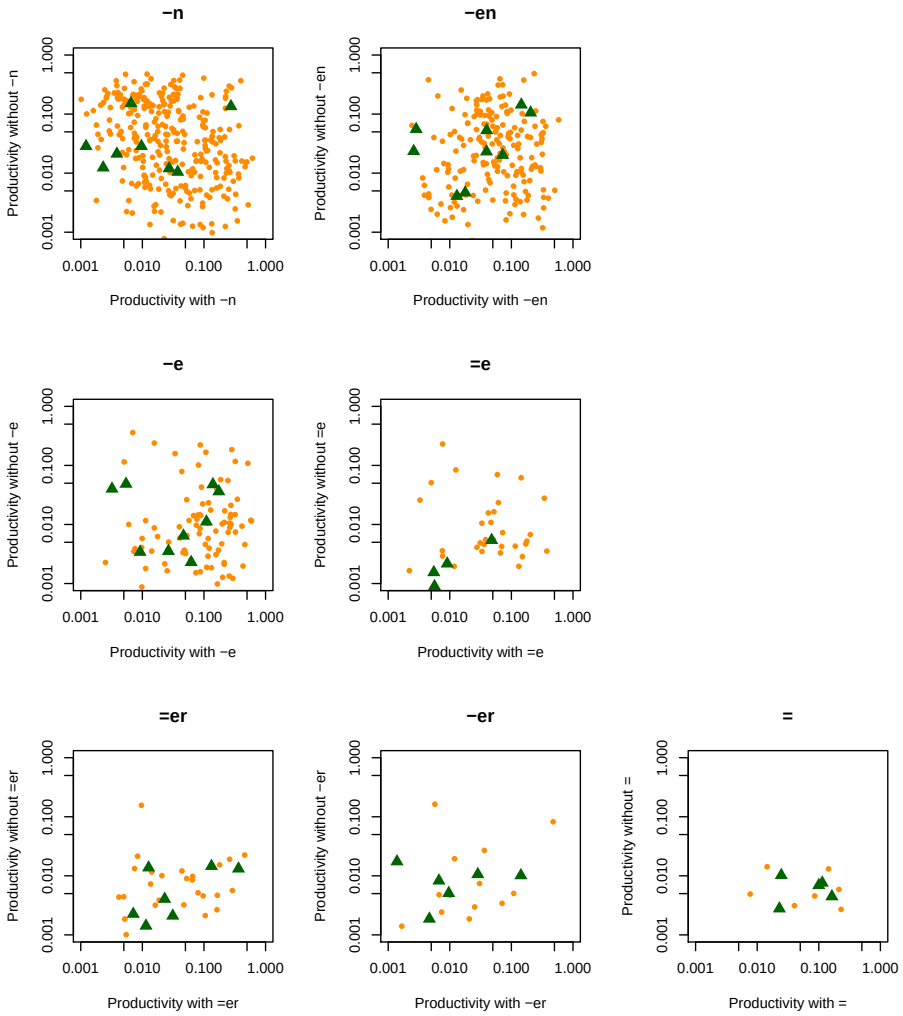

Fig. 2 Selection of the N1s for the corpus study (triangles) in the context of all possible N1 candidates meeting the minimal criteria (dots); the x-axis and y-axis plot the potential productivity of N1 with PL and NPL, respecitvely; axes are on a logarithmic scale

Annotating singular and plural was unproblematic, except where singular and plural forms were homographic *and* the context did not help to disambiguate the two. Such cases were discarded and not used for the study.[19]

Determining whether the semantics of N2 and the semantic relation holding between N1 and N2 forces N1 to have plural semantics was more intricate, and we found several classes

[19] Since this part of the study is crucial and strongly related to the scientific inferences we are going to draw, we did not rely on the automatic annotation for plural (and other morphological categories) available in DECOW16A. The basic annotations were made by the second author of this paper and a student assistant, but they were *all* cross-checked by the first author in order to minimise the amount of human annotation errors.

of N2s characteristic of this relation. A clear case of plural-inducing N2s are collectives such as *Gruppe* 'group' as in *Kind-er+gruppe* or *Kind-gruppe* 'group of children' or *Haufen* 'heap/pile' as in *Brett-er+haufen* or *Brett-haufen* 'pile of boards'. Even metaphorical collectives are usually unproblematic, for example *Regen* 'rain' in compounds like *Zitat-e+regen* or *Zitat-regen* 'rain of quotations'. Another clearly discernible group are reciprocals such as *Wechsel* 'swap/exchange' as in *Räd=er+wechsel* or *Rad+wechsel*, 'swapping of tyres'. Furthermore, there are N2s such as *Distanz* 'distance' as in *Löch=er+distanz* or *Loch-distanz* 'distance between (the) holes', which were annotated as plural-inducing N2s if it became clear from the context that a distance *between* several objects was referenced. Also, compounds with N2s like *System* 'system' as in *Element-e+system* or *Element-system* 'system of elements/periodic system' were annotated as containing an N1 with a forced plural interpretation if the reading was clearly that of a 'system of (several) elements'.

In addition to these fairly clear-cut cases, there was a second group of compounds in which the plural-inducing quality of the N2 was strongly dependent on context and world knowledge. Most prominent among these are N2s which denote a container of some sort. Examples include *Äpfel=+lager* or *Apfel-lager* 'storage for apples', *Brief-e+katalog* or *Brief-katalog* 'catalogue of letters', and *Lied-er+buch* or *Lied+buch* 'book of songs/songbook'. In theory, the N2s in these compounds *could* denote some sort of container which holds only one object (for instance, it is conceivable – if unlikely – that the storage space for apples could have only one apple in it), but both world knowledge and the particular context in which the compound appears render this sort of interpretation impossible. However, since the interpretation of this kind of containment compound is very context- and world-knowledge-dependent, we refer to these as extended cases and will analyse them separately from the clear-cut cases shown above. The presentation of the final sample will show the results for only the first, clear-cut cases under the label of *strict annotation* and the combined results for both the clear-cut and the extended cases under the label of *lax annotation*.[20]

4.3 Results

As mentioned in Section 4.2, the size of the manually annotated sample of 48 N1s was $n = 9,414$.[21] All exemplars were annotated for internal and external plural, and we treat the two annotations as two sub-studies, starting with the external plural sub-study.

There are 6,367 singular compounds and 3,048 plural compounds in the sample. Also, there are 4,708 cases with PL and 4,707 with NPL. The question is whether there is an

[20] We think it is encouraging that Banga et al (2013b) found the stable and expected results even using a more cavalier operationalisation of what they call conceptual plurals. "These form types could be divided into two conceptual types: the modifier of the compound was either conceptually singular (e. g., *bananenschil* 'banana skin' and *ballonvaart* 'balloon ride') or conceptually plural (e. g., *aardbeienjam* 'strawberry jam' and *appeltaart* 'apple pie')". The problem with *strawberry jam* is that *strawberry* could address the generic meaning ('strawberry type') or a mass noun version ('strawberry substance'), in which case the numerosity of referents of N1 would be much weaker conceptually. We had to assume that there would be more confounding factors in the corpus study than in a controlled experiment. Therefore, to be on the safe side, we did not consider hazy cases like these to be internal plurals.

[21] In the recent years and decades, practitioners, statisticians, and philosophers of science have been divided over the essential question of how to analyse quantitative data. Some have even been proclaiming a *statistical crisis* to both wider audiences (Gelman and Loken 2014) and to more restricted communities of practitioners (Gelman and Geurts 2017). We feel that corpus linguists have not yet picked up on the debate fully, with the effect that most papers do not make explicit which statistical philosophy they follow. We adopt the statistical philosophy of Ronald A. Fisher, and readers not familiar with the debate or the different statistical traditions are encouraged to read Appendix A.

considerably high number of exemplars with PL in a plural compound such that we might conclude that there is an external plural effect. The contingency table showing the two variables' bivariate distribution is given in Table 3.

	Pl. compound	Sg. Compound
NPL	1,574	3,133
PL	1,474	3,234

Table 3 Contingency table for external plural sub-study

A χ^2 test on the contingency table produces a significant result at sig $= 0.05$ ($\chi^2 = 4.786$, $df = 1$, $p = 0.029$). However, despite the very large sample size, the p-value is relatively close to 0.05, and the effect size is quite low at $v = 0.023$, which indicates that the effect is probably spurious. To make matters worse, an inspection of the (standardised) residuals in Table 4 shows that the spurious effect even has the wrong direction, i. e., PL is less frequent in plural compounds than expected under the null.

	Pl. compound	Sg. Compound
NPL	2.210	-2.210
PL	-2.210	2.210

Table 4 Standardised residuals from χ^2 test of external plural sub-study

Thus, we have found no evidence to support the external plural hypothesis in the global analysis of the data set.[22] There might, however, be item-specific differences between N1s or groups of N1s which take different PLs. In order to check for this, we calculated one χ^2 test for each N1 with an approximate sample size of $n = 200$ for each test. Since we are testing a family of connected hypotheses, p-values need to be corrected (correction for group-wise error). We used Šidák's method (Šidák 1967), which is slightly less conservative than the well-known Bonferroni correction (at least for uncorrelated tests).[23] The low counts in some cells in our data make the χ^2 approximation inexact, so we used a Monte Carlo replacement for the traditional χ^2 test (Hope 1968) as implemented in the standard `chisq.test` function in R (R Core Team 2014). The procedure calculates these marginal sums and then generates random permutations for the contingency table given the marginals. We used $b = 10,000$ permutations. Finally, the effect sizes (v) were calculated and multiplied by the sign of the top left cell of the residual table. This means that, in addition to quantifying the magnitude of the effect size, they also reflect whether, for each N1, the co-occurrence of PL and plural semantics on N1 is preferred (positive sign) or dispreferred (negative sign).

These results confirm those of the global test reported above. Only one test reaches sig $= 0.05$. The effect sizes are centred around 0 with a mean of $\bar{v} = -0.022$, a median of

[22] Which is not the same as finding evidence *against* it.

[23] We also tried Holm's and Hochberg's methods, but the differences between all those methods are small enough that they do not lead to different interpretations of the results. Instead of adapting the sig level, we corrected the p-value directly using $p_S = 1-(1-p)^m$ where m is the number of tests ($m = 48$ in this case).

$\tilde{v} = -0.024$, and a standard deviation of $s(v) = 0.116$. This is expected in a situation where there are no effects.

The picture changes when we move on to the analysis of the internal plural sub-study. Here, we examine whether the occurrence of PL is more frequent in internally plural compounds, i. e., compounds with an N2 that forces plural semantics on the N1. Recall from Section 4.2 that we carefully distinguished between clear cases and less clear cases, calling the two annotation schemes *strict* and *lax*. We therefore specify two values separated by a semicolon for the relevant statistics (*strict*;*lax*). The sample contains 911; 1,237 cases with an internal plural relation and 8,503; 8,177 cases with no internal plural relation. Table 5 shows the contingency table for the two variables PL vs. NPL and internal plural relation vs. no internal plural relation.

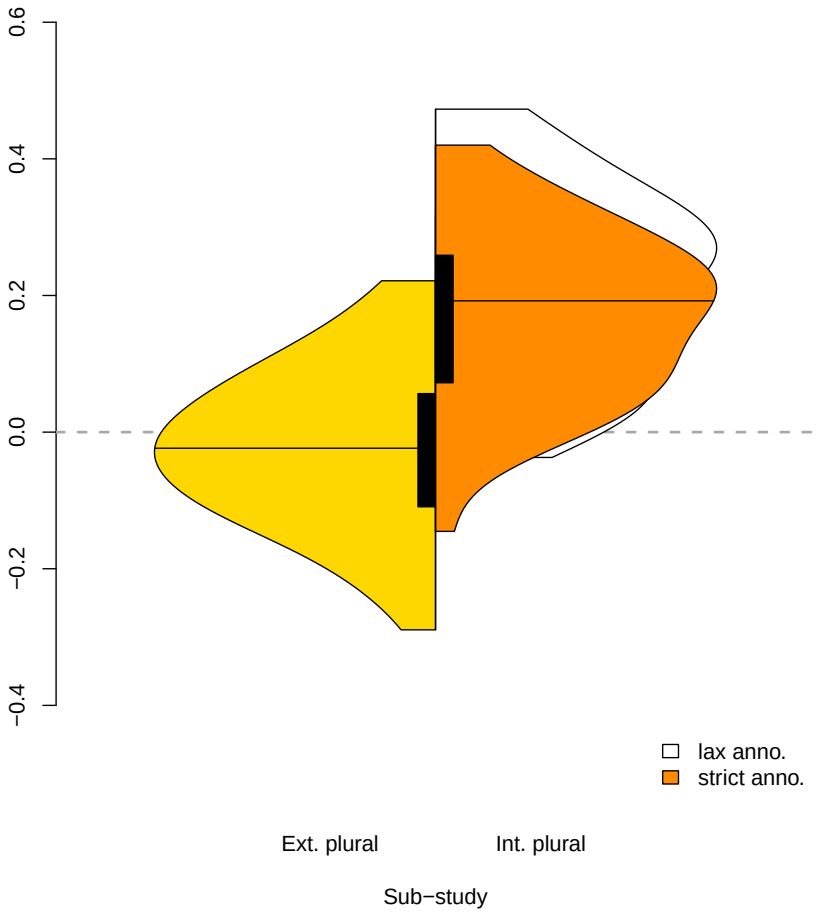

Fig. 3 Distribution of signed Cramérs *v* scores for the internal and external plural data

	Internal pl. relation	No internal pl. relation
NPL	214	4,493
PL	697	4,010

Table 5 Contingency table for internal plural sub-study

A χ^2 test reaches sig $= 0.05$ ($\chi^2 = 282.343; 448.259$, $df = 1; 1$, $p \approx 0; 0$). However, the overall effect size is still not very high at $v = 0.173; 0.218$. At least for the lax annotation scheme $v \geq 0.2$, making the effect worthy of mention. Individual (per-N1) tests were also calculated with the exact same parameters as described above for the external plural sub-study. Figure 3 compares the distribution of v values for the external plural sub-study (left) and the internal plural sub-study (right) in the form of a split violin plot.[24] For the internal plural sub-study, strict and lax annotation are both shown. Obviously, the difference between the two annotation schemes is not huge, and virtually identical conclusions can be drawn based on either of them. While the external plural has zero average effect, the internal plural has an average effect of roughly 0.2.

Even more informative is a closer look at the individual results. Figure 4 represents for each N1 the signed effect strength in the internal plural sub-study as the dot's location on the horizontal dimension and the p-value through the colour coding. The darker the dot is, the lower the p-value. The colour mapping is logarithmic, such that any dot representing a p-value above sig $= 0.05$ is already significantly lighter than the darkest colour. The N1s are arranged in groups defined by their PL. The groups are sorted by their mean effect strengths, and within each group N1s are sorted by their effect strengths.

The detailed analysis shows that there are considerable differences between LEs. N1s with =*e* and =*er* especially show good effect strengths as well as reaching reasonable significance levels. Those with -*er* and = show mixed results, and for -*e*, -*en*, and -*n*, we find only very weak effects and non-significant p-values, which would normally be dismissed completely. However, even though the results for the last mentioned PLs are negligible by themselves, it should still be noticed that all N1s show a positive effect except *Ei* 'egg' and *Katze* 'cat', which lean ever so slightly towards the negative side.

This strong positive trend is not at all the distribution of v scores which one would expect if there were no general co-occurrence preference between PL and an internal plural relation. Rather, roughly as many negative as positive v scores distributed around 0 would be expected, as was the case in the external plural condition; see Figure 3. These results thus support the hypothesis that there is a general co-occurrence preference between PL and an internal plural relation.

Finally, it is noteworthy that the order of the PLs in Figure 4 is the mirror image of their order in Figure 1 and Figure 2. The more marked and thus the less type-frequent a plural marker is, the stronger it tends to be interpreted as a plural marker even when used as a LE. We will return to all of these results in our theoretical interpretation in Section 6. But first, we report the results of a split-100 experiment in Section 5.

[24] Violin plots are extensions of standard box plots; see Hintze and Nelson (1998). The horizontal line represents the median and the black vertical bar the interquartile range (i. e., the range of the middle 50% of the sample). The additional outer shape represents a kernel density estimate of the sample.

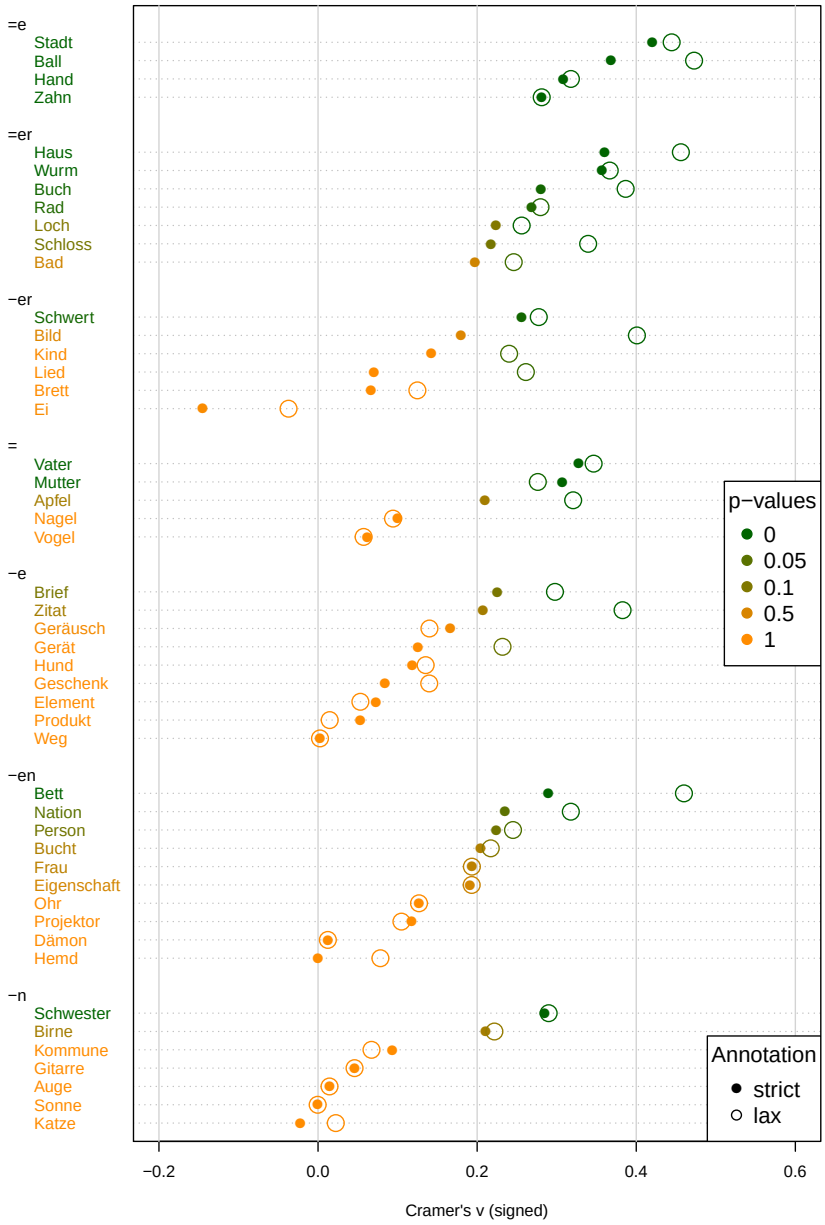

Fig. 4 Individual per-N1 effect strengths and Šidák-corrected p-values for the internal plural sub-study; groups (PLs) are sorted by descending overall effect strength (strict); per group, N1s are sorted by descending effect strength (strict)

5 Split-100 experiment

5.1 Design, choice of stimuli, and participants

This experiment was designed to examine how strongly native speakers of German prefer a PL through two sub-experiments investigating internal and external plural relations. A split-100 task was chosen because it is claimed that participants make subtler judgements compared to a binary forced-choice task (Ford and Bresnan 2013, Verhoeven and Temme 2017, to appear). In a split-100 task, subjects are offered two options (here: a compound with PL or NPL) and they can weigh their preference for either of them, assigning integer values between 100;0 (clear preference for option one) and 0;100 (clear preference for option two) to the tuple of options.

The experiment was conducted using PsychoPy (Peirce 2007) and contained two sub-experiments merged into one run for each subject. Subjects made eight decisions pertaining to the internal plural sub-experiment and eight decisions pertaining to the external plural sub-experiment. In addition to these 16 targets, an experiment run contained 41 fillers, which results in a target-to-filler ratio of 1:2.5. A training phase with five sentences/decisions preceded the experiment.

Subjects were presented with sentences containing a blank where a compound should go. The two variants of the compound (with PL and NPL) were shown on the same screen below the sentence with a slider in between, which could be moved freely to assign a preference between 100;0 and 0;100. The corresponding numbers were displayed dynamically as subjects moved the slider. When they were satisfied with their decision, subjects pressed a button to store it and continue on. After each sentence, subjects answered a simple distractor question by pressing one of the digits 1–9. The questions were not related to the sentences and were simple arithmetic or counting exercises. Finally, it should be noted that the order of the stimuli and fillers was randomised for each participant.

The choice of stimuli was guided by the exploratory data analysis described in Section 3. First of all, we chose N1s which have roughly equal productivity with PL and NPL. We then tried to find for each N1 some semantically appropriate N2s which clearly trigger internal plural semantics and ones which clearly do not. Finally, we checked the frequencies of the resulting N1+N2 compounds with PL and NPL in the corpus, because we wanted to use compounds as stimuli which were productively formed for the subjects inasmuch as they had never used or heard/read them before (at least with a high probability). The process turned out to be an iterative one because compounds meeting all desired criteria were difficult to find. Table 6 presents the results of the selection process. The table shows the potential productivity P^p for the N1 with PL and NPL. These roughly match in many cases, for example $P^p_{PL}(\text{Bad}) = 0.013$ and $P^p_{NPL}(\text{Bad}) = 0.014$. The frequency classes c of the full compound with PL and NPL is also given, where class 28 corresponds to a token frequency of 1, and no frequency class (–) is assigned to words with a token frequency of 0. Additionally, the total difference in token frequency of the two compounds is specified (Δ_f), where the maximum absolute difference is 63, which is clearly not extreme, given that the corpus contains 21 billion tokens and approximately 15 billion words.

The first eight compounds shown in the table were used as stimuli for the internal plural sub-experiment. The N2s *Kooperation* ‘cooperation’, *Zusammenlegung* ‘merger/unification’, *Bündel* ‘bundle’, and *Sammlung* ‘collection’ are clear triggers of internal plural. In contrast, *Eingang* ‘entry’, *Beschriftung* ‘label(ling)’, *Schliff* ‘cut/sharpening’, as well as *Abdruck* ‘mark/impression (physical)’ cannot trigger internal plural semantics at all. Because it is sensitive to the specific semantic relation between N1 and N2, the internal plural sub-

N1	LE	N2	P^p_{PL}	P^p_{NPL}	c_{PL}	c_{NPL}	Δ_f	Gloss
Bad	=er	Kooperation	0.013	0.014	25	28	6	bath cooperation
Bad	=er	Eingang	0.013	0.014	28	23	-27	bathroom door
Weg	-e	Zusammenlegung	0.009	0.003	–	–	0	path merger
Weg	-e	Beschriftung	0.009	0.003	–	26	-3	pathway label
Brett	-er	Bündel	0.029	0.011	28	–	1	board bundle
Brett	-er	Schliff	0.029	0.011	–	28	-1	board cut
Schwert	-er	Abdruck	0.143	0.010	–	28	-1	sword mark
Schwert	-er	Sammlung	0.143	0.010	25	26	5	sword collection
Haus	=er	Abbruch	0.011	0.001	23	22	-52	building demolition
Kraft	=e	Stärke	0.008	0.003	26	23	-20	force strength
Grab	=er	Buchung	0.016	0.003	–	–	0	grave reservation
Punkt	-e	Farbe	0.008	0.004	–	23	-38	dot colour
Blatt	=er	Analyse	0.014	0.007	–	22	-63	leaf analysis
Bett	-en	Länge	0.018	0.005	21	21	-30	bed length
Last	-en	Berechnung	0.010	0.006	22	22	25	load calculation
Hemd	-en	Schlitz	0.072	0.020	28	25	-8	shirt slit

Table 6 Selection of stimuli for split-100 experiment; P^p_{PL} and P^p_{NPL} are the potential productivities of the N1 with PL and NPL, respectively; c_{PL} and c_{NPL} are the frequency bands of the compound with PL and NPL, respectively; Δ_f is the difference in raw token frequency between the compound with PL and NPL

experiment might be sensitive item-specific effects pertaining both to concrete combinations of N1 and N2. In order to control for such effects, we chose only four different N1s and combined each with an N2 triggering internal plural and one not triggering it. Given the four N1s, combining internal-plural N2s and non-internal-plural N2s with PL and NPL created sixteen targets for the initial plural sub-experiment alone. So that participants were not exposed to all sixteen of these, we randomised the stimuli and participants saw either the version with PL or with NPL of any one N1+N2 combination. For the external plural sub-experiment, we only used N2s which could not trigger internal plural meaning, and we therefore decided that increased variety of N1s was better and used eight different N1s.

Participants declared themselves to be native speakers of German with no reading disorders. They were recruited in first-semester linguistics classes at our university during the first four weeks of the summer 2017 term. They were all majoring in German Language and Literature but had not yet had a university-level introduction to linguistics. We had 31 participants. 24 participants declared themselves to be female, seven to be male. Age varied between 19 and 31 with a median of 21.

5.2 Results

The results are clearly in line with the findings from the corpus study reported in Section 4. Figure 5 shows the distribution of the split-100 responses in the form of violin plots. Notice that all ratings were re-mapped such that 0 represents a clear preference for NPL and 100 a clear preference for PL, although in the actual experiment, the assignment of 0 and 100 to the conditions was randomised.

The external plural sub-experiment (right panel of Figure 5) clearly had a negative result. For both singulars and plurals, subjects strongly prefer a NPL (for singular compounds, the median and mean rating are 15.500 and 35.056, and for plural compounds, they are 20.000 and 28.169).

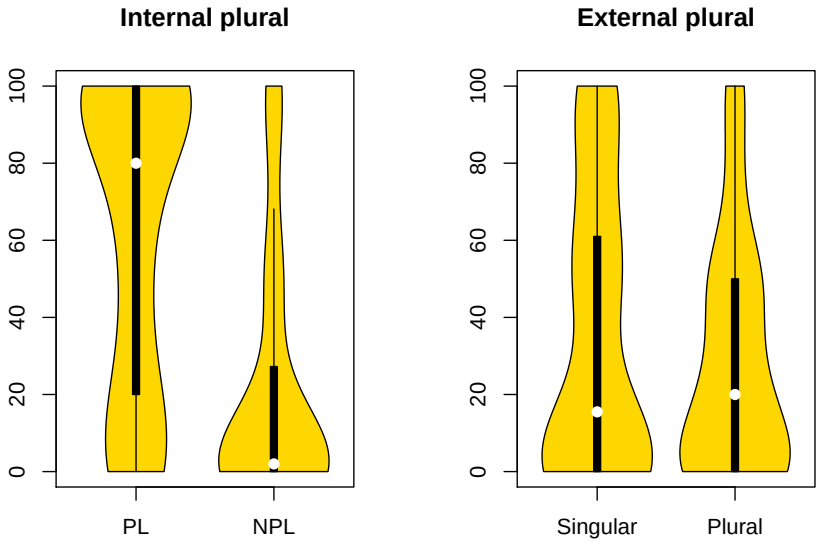

Fig. 5 Distribution of Split-100 responses in both sub-experiments by conditions

The internal plural sub-experiment (left panel of Figure 5) had a clearly positive outcome. The median and mean with NPL are 2.000 and 20.782, but with PL, they are 80.000 and 61.742). Even a simple exploratory data analysis thus shows that a PL strongly co-occurs with forced plural semantics on N1.

Analysing split-100 experiments with inferential tools such as generalised linear mixed models (GLMMs) is not as straightforward as Ford and Bresnan (2013) seem to suggest. First, the dependent variable is something like a proportion or a percentage and cannot be treated as a count or a numeric variable for modelling purposes, because assumptions underlying the modelling algorithms (such as homogeneity of variance) would be violated. For proportions, beta regression can be used. However, subjects tend to assign ratings of 0 and 100 very often (see Figure 6), which leads to so-called *zero inflation* and *one inflation* (see Zuur et al 2009 for a detailed account of practical modelling with zero inflation targeted at practitioners). Since in the case at hand, we should also account for subject-level variation by including a random effect, the appropriate model is a *zero- and one-inflated mixed beta model*. The fact that control for subject-level preferences is important is illustrated by Figure 6, where individual preferences in the internal plural sub-experiment are shown per condition.

A package for R which implements zero- and one-inflated mixed beta models is `gamlss` (Rigby and Stasinopoulos 2005). It is actually suitable for fitting general additive mixed models (GAMMs), but GLMMs can be fitted as a special case of a GAMM. In this case, no smoothing function is estimated for the fixed parameters, and a special smoother (`random()` in `gamlss`) is used for grouping factors to be used as random effects. We estimated the parameters of a GAMM specified as `Rating ~ Condition + random(Subject)` based on the data from the internal plural sub-experiment with the `BEINF` family for inflated beta

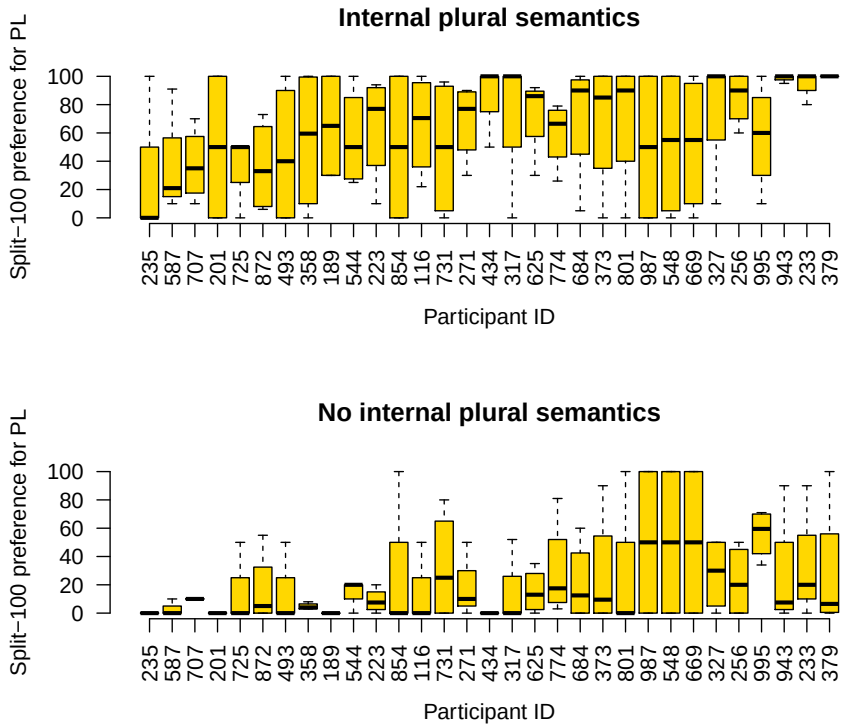

Fig. 6 Boxplots of the distribution of individual (per-subject) split-100 responses in the internal plural subexperiment by conditions; subjects were ordered from left to right by their mean rating across both conditions

models ($n = 248$).[25] Ratings had to be mapped from $[0,100]$ to $[0,1]$. The only parameter of interest in the current setting is the coefficient for the `Condition` variable, which is estimated at -0.893 with an estimated standard error of 0.195. It reaches $sig = 0.05$ ($t = -4.585$, $p \approx 0$). The model has an AIC of 517.008. In comparison, a model without the predictor of interest (`Rating ~ random(Subject)`) has an AIC of 531.801, which is worse by 14.793. In other words, the model corroborates the descriptive analysis in Figure 5 using advanced inferential tools.

In the present experiment, subjects showed very clear reactions to different potential sources of plural semantics. The external plural condition did not trigger PLs more than NPLs, but there is solid evidence that speakers favour a PL in compounds where an internal plural relation holds between N1 and N2. In Section 6 we will discuss the theoretical implications of this result together with that of the corpus study reported in Section 4.

[25] Technically, the given formula is the formula for the μ parameter.

6 Conclusion

Both in the corpus study (Section 4) and the split-100 experiment (Section 5), we found evidence that writers prefer to use a PL when N1 necessarily has a plural interpretation due to compound-internal semantics (internal plural effect). We found no evidence for a similar influence of a plural suffix on the whole compound (external plural effect). This effect could either have been semantically motivated in some cases or be a simple compound-internal number agreement tendency, neither of which appear to be the case. In light of the suspicions raised in the literature that there might be plural effects within German compounds, as well as the previous research on Dutch LEs (see Section 2.2.1), our results do not come as a surprise. Regardless of the diachronic development of LEs, writers use PLs as cues to the interpretation of N1+N2 compounds (see also Banga et al 2013b, 212). The form of an N1 with a PL is one which writers associate strongly with semantic plurals, and it appears that they do so also when that form occurs in a compound. A detailed look at our results reveals more about the actual mechanisms at work.

First, the effect strengths in the corpus study were weak to intermediate, and while the observed effect was clearer in the split-100 experiment, we are nowhere near a categorical split. There is also substantial per-speaker variation and uncertainty. (Readers can revisit Figure 5 and Figure 6 to convince themselves that this is the case.) So, we have to answer the question of why the overall effect is not stronger. As pointed out in Section 2.2.1, Banga et al (2013a, 45) suggested that there may be a general tendency to avoid LEs, based on the finding by Libben et al (2002) that compounds with LEs come with higher processing costs. This could account for the high percentage of zero linkages (see Section 3.2, especially Table 2). Furthermore, regardless of whether one favours a rule-based or a similarity/exemplar-based view (see Section 2.1.2), in many cases the use of a PL is highly unlikely for unrelated reasons such class membership or phonotactics. A clear case are words ending in full vowels, which have an *-s* plural but resist *-s* linkages strongly. For example, while *Auto-s+kollektion* 'car collection' cannot be excluded with absolute certainty, *Autokollektion* would be preferred by a huge margin. Such factors stand in the way of establishing a stronger link between PLs and plurality.

Second, we must ask why the corpus-based results were weaker than the split-100 results. We propose that this is due to the fact that we made a considerable effort to ensure that the stimuli in the experiment consisted of compounds which were most likely novel to the subjects (see Section 5.1). In creating the samples for corpus study, however, we did not differentiate between compounds strongly established in language use and novel compounds. Therefore, the corpus sample contains lexicalised compounds with conventionalised idiosyncratic linkages. Of course, we would expect that even lexicalised compounds follow the tendency to associate PLs with plural meaning at least to some extent. Thus, simply filtering compounds which have a high token frequency from concordances could obscure part of the picture and may not be in our best interests when trying to study this phenomenon. More theoretical, empirical, and even methodological work on this is clearly required.

Third, we must ask why PLs differ with respect to their tendency to be used to mark plurality in the corpus study.[26] The order of preference established in Section 4 (especially Figure 4) was *=e* ≫ *=er* ≫ *-er* ≫ *=* ≫ *-e* ≫ *-en* ≫ *-n*. The strength of the internal plural effect is thus by and large negatively correlated with the type frequency of nouns which take the respective plural (see Section 3.2, especially Table 2, and Section 4.1, especially Fig-

[26] The number of possible stimuli in the experiment was too low due to constraints imposed by the experimental procedure to allow a more differentiated look at variation between PLs.

ure 2). Additionally, the inflectional suffixes *-en* and *-n* occur in many positions in German nominal inflection and therefore provide the lowest cue validity for plurality. For example, there is a dedicated dative marker *-(e)n* attaching to all plurals except *-(e)s*, where it is excluded for phonotactic reasons. Also, weak nouns mark all non-nominative singular forms with *-(e)n*. Finally, the so-called *weak* and *mixed* adjectives end in *-en* in all oblique (dative and genitive) singular forms and the masculine accusative singular. On the other hand, *-er* is much more exclusively a plural marker (with some exceptions in the so-called *strong* adjectival paradigm), and umlaut is clearly reserved to mark plural in German nominal inflection.[27] In short, stem umlaut is a highly salient cue, whereas *-e* (schwa) and *-(e)n* (usually realised as a syllabic nasal) are the least salient of plural markers. Taken together, this means that plural interpretation of PLs is stronger with what Köpcke (1993) calls a higher *Signalstärke* 'signal strength' of the PL as a plural marker. His criteria for high signal strength are are high salience, high validity, low type frequency, and a high degree iconicity.[28] We conclude that writers associate plural meaning more strongly with PLs if the PL itself is more strongly associated with plurality, i. e., has greater signal strength. Furthermore, we consider it likely that the correlation between type frequency in our sense and the increased tendency for plural interpretation of PLs is merely accidental. The more salient and cue-valid plural patterns happen to be the less type-frequent ones for historical reasons.

Fourth, an explanation of why only the internal plural effect was observed is necessary. In Section 1, we argued that restrictive frameworks tend to elevate universal tendencies to hard universal constraints. Among the proposed morphological constraints were those requiring that there be no inflection inside products of word formation. We would like to point out that the tendencies that we have uncovered concern the internally licensed inflectional category of number only (see Section 2.2.1). The major externally licensed nominal inflectional category, namely case, would still be highly unlikely to occur between N1 and N2, obeying the putatively universal constraints. We suggest that this follows from the nature of compounds. Case could conceivably only be used to specify the grammatical relation between a compound's constituents, and this is something not usually found in compounds, but rather in complex noun phrases. Number, on the other hand, has nothing to do with the grammatical relation between the constituents. Writers use the optional plural marking when the individual and joint conceptual representation of the constituents allows this, but it has nothing to do with syntagmatic grammatical relations. However, the external plural effect would require that a formed compound be somehow transparent for grammatical categories attaching to its whole. This is atypical of compounds.

Fifth, we would like to propose an explanation of why our findings seemingly contradict those of Koester et al (2004). They show in an experiment using event-related potentials (ERP) that mismatches between the semantic (non-)plurality of N1 and PLs/NPLs do not lead to the expected N400 effect, concluding that PLs and NPLs are not used by hearers to decode the conceptual structure of N1+N2 compounds. In contrast to the ERP experiment, we only considered the production side of the effect by looking at usage data from a corpus and preferences in a production-oriented decision task. Since plurality is only marked optionally within compounds, as we have argued extensively above, hearers cannot and need

[27] Some comparative and superlative adjectives take an umlaut in addition to the corresponding suffixes, but the formation of comparatives is clearly something very different from case and number inflection. Also, case and number inflection attaches additionally to the right of the comparative suffix.

[28] We do not discuss the more complicated notion of iconicity here for reasons of space. By type frequency, Köpcke means the number of different forms in a paradigm, and what we have described as cue validity above encompasses Köpcke's validity and type frequency. For us, type frequency is rather the number of different nouns with which a plural marker occurs.

not rely on PLs as plural markers. However, the fact that hearers do not rely on plural marking does not necessarily prevent writers from using plural forms where a plural meaning is intended. This, then, raises the question of whether the plural is really *marked* or just redundantly indicated by a PL. Only further research can shed light on such questions.[29]

Sixth and finally, Banga et al (2013a) found that German L2 learners of Dutch do not react as strongly to the Dutch PL *-en* as Dutch L1 speakers. This could be due to the fact that, as we have shown and explained, Germans attribute the lowest plural interpretability to *-en* in German, rather than due to an across-the-board weakness of the relationship between plurality and PLs in German, as the authors speculated (Banga et al 2013a, 45). If there really is an interference effect from the German L1, then the Dutch marker being accidentally identical to the German marker with the lowest signal strength would naturally lead to low affinity for a plural interpretation in compounds. For Dutch speakers, however, given the complete absence of nominal case inflection and much reduced system of plural markers, differences in signal strength are irrelevant.

Our study has helped to show that German PLs can indeed have a plural interpretation. More importantly, we have shown that there are differences in when this interpretation is available. We have also provided cognitively oriented explanations for these differences. We see many possible routes for future research on the subject. For example, more refined corpus studies taking into account the degree to which a compound is novel or established should be made. Also, the differences between the production and the reception side of plurality in compounds need further research, especially considering that for each side, only one study has been conducted (namely Koester et al 2004 and this one). While there appears to be a solid connection between PLs and plurality, this connection's exact nature still remains to be characterised fully.

[29] Some relatively simple experiments could provide clarification about the reception of PLs. These could involve subjects associating objects or illustrations with truly ambiguous compounds such as *Apfel+teller* 'apple plate' and *Äpfel=+teller*. For more implicit testing, the task might be set up in the visual world paradigm.

Acknowledgements

We would like to thank (in alphabetical order) Jennifer Dailey-O'Cain, Matthias B. Döring, Jordan Lachler, and Ulrike Sayatz for valuable discussions and comments. Furthermore, we thank Ulrike for helping us recruit the participants for the experiment. Luise Rißmann did sterling work helping with the annotation of the concordances and supervising a majority of the experiments. We are, of course, fully responsible for all residual inadequacies, errors, and omissions.

Compliance with ethical standards

The split-100 rating experiment reported in this paper was completely anonymised. Participants signed standard consent forms and were given the option of asking questions about the nature of the experiment before and after they took part in it. They were granted the right to revoke their signed declaration of consent (through anonymised personal codes), which none of them did. Participants received no payment, but they were awarded credit in partial fulfillment of the requirements for a first-semester class in German linguistics at Freie Universität Berlin. The participants could specify their gender on a voluntary basis and they were given no pre-formulated options.

A On statistical analyses

In Section 2, we derived two operationalisable working hypotheses from our main substantive hypothesis that there is a connection between pluralic linkages and plural interpretations of N1. In this appendix, we explain our position on data analysis and so-called *hypothesis testing*. The most widely used statistical system is *Null Hypothesis Significance Testing* (NHST), and it is one of the *frequentist* systems of statistical inference. In NHST, researchers attempt to substantiate the existence of an effect (such as a positive connection between plural semantics and pluralic linking elements) which is predicted to exist by their favoured theory by means of conducting an experiment in which the effect is measured. Then, the probability p (the so-called *p-value*) of obtaining the observed measurements or more extreme measurements under the assumption that there is actually *no* effect (the *null hypothesis* or just the *null*) is calculated. If this probability is lower than a certain threshold (usually called the α*-level*), the null hypothesis is *rejected*, which is taken as evidence that the hypothesis derived from the theory is correct. It is often incorrectly stated that "the experiment/test shows that the probability that the null is correct is p" or "is lower than α". This approach is riddled with philosophical and statistical problems and has led to the promotion of of bad scientific practice. Among the most ardent critics are Gigerenzer (2004), Colquhoun (2014), and Munafò et al (2017), and the editors of the journal *Basic and Applied Social Psychology* have even banned the use of p-values in an actionist attempt to tackle problems of bad science related to NHST (Trafimow and Marks 2015). Critics often propose to abandon frequentist inference altogether and adopt a Bayesian approach, which itself is not without philosophical and practical problems (see, for example, Mayo 1996, Senn 2011). Other have proposed abandoning statistical inference proper in favour of confidence intervals and effect sizes (Cumming 2014), sometimes not noticing that NHST confidence intervals are not considerably different from NHST p-values, as Perezgonzalez (2015a) shows in reply to Cumming (2014).

However, there is no need to abandon frequentist inference or p-values simply because they have been abused. A great many statisticians and researchers have shown that the major problem with NHST is that it is a mixture of the statistical philosophies of Ronald A. Fisher on the one hand and Jerzy Neyman and Egon Pearson on the other hand (see Goodman 2008, Perezgonzalez 2014, Perezgonzalez 2015b, Greenland et al 2016; see also Lehmann 1993 and Lehmann 2011 for an overview of these two philosophies and the history of their development). We follow Fisher's statistical philosophy, and we briefly compare it to Neyman and Pearson's now.

Neyman and Pearson developed a system where two hypotheses are specified: the *main hypothesis* ($\mathrm{H_M}$) and the *alternative hypothesis* ($\mathrm{H_A}$), and they have to exhaust the probability space ($p(H_M \cup H_A) = 1$). The goal is to accept either of these hypotheses and reject the other, where typically $\mathrm{H_M}$ is the hypothesis predicted by the experimenter's favoured theory and the one they would like to accept. The reason why the Neyman-Pearson approach can be hard to implement is that $\mathrm{H_M}$ needs to be specified *precisely*, i. e., including the effect size. For example, if the experiment is a reading time experiment contrasting reading times under two distinct conditions, then the expected increase in reading times needs to be specified numerically. If this is possible, then researchers can calculate the risk of incorrectly accepting $\mathrm{H_M}$ when it is false (α) and the risk of incorrectly accepting $\mathrm{H_A}$ when it is false (β) *given specific sample sizes*, then setting the optimal sample size and choosing the optimal test procedure. Especially Neyman designed this system explicitly with the idea in mind that researchers end up doing the right thing in $1 - \alpha$ of all cases if they follow this protocol. No inference with respect to the ultimate truth of a specific hypothesis at hand was ever intended by Neyman (see Neyman and Pearson 1933, 290–291 and Neyman 1937, 349 for very clear statements to this effect). In empirical linguistics (both corpus-based and experimental), following the Neyman-Pearson protocol is often impossible because theories do not predict effect sizes.

Fisher developed a system where the probability of a specific outcome of a random experiment (or a more extreme outcome) *if there is no effect* (the $\mathrm{H_0}$ or *null hypothesis* or simply the *null*) is calculated as the p-value. It cannot be stressed enough that this is the probability of obtaining such results *before the experiment is conducted* and *taking into account the design of the experiment*.[30] Fisher (1926, 504) suggests an informal, adaptive, and approximate *threshold of significance* (or *sig*), for example 0.05, below which researcher might suspect that there is something going on (see also Section 4.4 in Lehmann 2011 for a detailed summary of Fisher's positions). While Fisher did not directly recommend the inspection of p-values, he recommended that experimenters set *sig* appropriately based on previous experimental or theoretical knowledge (see Chapter 4 of Lehmann 2011 and Perezgonzalez 2015b). Furthermore, p-values can be corrected after the experiment, for examples if many conceptually related tests are performed, which increases the actual error rates relative to the nominal ones. The most important pitfalls and misunderstandings (directly translating into some of the false assumptions common in NHST) in Fisher's framework are:

[30] It is *not* a Bayesian posterior probability which allegedly quantifies the credibility of a hypothesis given the data.

1. Researchers take a significant result as a proof of something, usually the hypothesised effect. In fact, significance only shows that either the null does not describe the actual world very well *or a rare event has occurred*. There is no way of knowing with any specifiably accuracy which of these is the case.
2. Practitioners take point (1) even further and make an inference from a single significant result to some substantive hypothesis such as "my whole theory is correct".
3. Researchers assign high importance to some significant result although the data only suggest that the null might be rejected, but that the effect is rather small.
4. If one runs a series of experiments and performs the corresponding tests in which the nulls are conceptually related, the actual probabilities of just a rare event happening increase, and each *p* or the *sig* level are too optimistic if left uncorrected.
5. Finally, practitioners might not have conducted a proper random experiment (wilfully or out of ignorance), thus changing the sample space and invalidating the actual computations.

Points (1) and (2) can be remedied by researchers being aware of the actual (low) importance which can be attributed to a single significant result. Furthermore, good use of previous experimental and theoretical knowledge in evaluating the actual p-values (although Fisher himself was not much interested in interpreting p-values) helps to make the Fisher approach more sound in practice. It also helps to do replications and perform meta-analyses. Problems with point (3) are easily avoided by looking at effect sizes (which are usually associated with the Neyman-Pearson approach but imported easily into a Fisherian procedure). Demanding that researchers should pay more attention to effect sizes is really just another way of saying that they should do proper exploratory/descriptive analysis of their data sets. Point (4) can be dealt with by applying corrections for group-wise error (which should not be called "α-level correction" under Fisher's approach, even if the two are mathematically equivalent).

Point (5) poses the most serious threat to validity in corpus linguistics. Fisher's logic of experimental design presupposes that test subjects were randomly chosen from the population of interest and that confounding factors are thus distributed randomly (see Chapter 2 of Maxwell and Delaney 2004 for a convenient overview). Experimentation in corpus linguistics but also in linguistics in general is marred by the fact that researchers often cannot specify their population of interest with high precision. The traditional discussion of the *representativeness* of a corpus does not help because it is more often than not centred around the concept of a corpus being "representative of a language" (as a whole), using as points of reference: (i) the distribution of texts or text types in the output of all speakers of a language (production-based), (ii) the distribution of the relevance of texts or text types in the whole speech community (relevance-based), or (iii) the distribution of speakers' exposure to different texts or text types (perception-based).[31] In Stefanowitsch and Flach (2016), a recent contribution where the perception-based view is argued to be valid in cognitively oriented corpus linguistics, a global view of representativity is addressed.[32] The population of interest has to be defined with regard to each experiment individually, and it might be something very specific (such as the written output of speakers of a certain age, in a specific register, etc.) instead of "the language" or "the average speaker" (across all communicative settings and modes). In social sciences, the concepts of *global and specific representativeness* (Bortz 2005, 86) are used to describe the relevant distinction.

If the population of interest cannot be specified precisely, as is often the case in corpus linguistics (including our study), and a global study is performed, then everything should be done to increase the validity of the study, prominently: (i) choose the most varied and large corpus available, (ii) regard the study as partly exploratory, even if statistical tests and previous theoretical knowledge (including predictions derived from theories) are used, (iii) be appropriately careful in the interpretation of the findings, ideally using additional sources of data such as experiments (see Bresnan et al 2007 for pioneering work in this area).

This is why we chose DECOW16A: it is very large but also contains a lot of variation (including non-standard writing). Also, as we demonstrated in Section 4, we consider the exploratory nature of our work more important than binary decisions of significance. We also do not claim that our results generalise beyond the type of texts contained in DECOW16A, especially not to spoken language. Furthermore, as was shown in Section 4, careful data analysis reveals more fine-grained effects than audacious global hypothesis testing. Finally, the experimental results reported in Section 5 are somewhat clearer than the corpus findings, which highlights the need to use corroborating evidence from several methods (see, for example, Arppe and Järvikivi 2007, Divjak et al 2016).

[31] For overviews from different perspectives, see Biber (1993), McEnery et al (2006), Leech (2007), Hunston (2008).

[32] "In this wider context, large, register-mixed corpora such as the British National Corpus [...] may not be perfect models of the linguistic experience of adult speakers, but they are reasonably close to the input of an idealized average member of the relevant speech community." (Stefanowitsch and Flach 2016, 104)

References

Anderson SR (1992) A-Morphous Morphology. Cambridge University Press, Cambridge

Arndt-Lappe S, Bell MJ, Schäfer M, Schlücker B (2016) Introduction: modelling compound properties. Morphology 26:105–108, DOI 10.1007/s11525-016-9285-4

Arppe A, Järvikivi J (2007) Every method counts: combining corpus-based and experimental evidence in the study of synonymy. Corpus Linguistics and Linguistic Theory 3(2):131–159, DOI 10.1515/cllt.2007.009

Augst G (1975) Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache. TBL Verlag, Tübingen

Baayen RH (2009) Corpus linguistics in morphology: morphological productivity. In: Lüdeling A, Kytö M (eds) Corpus linguistics. An international handbook, Mouton De Gruyter, Berlin, pp 900–919, DOI 10.1515/9783110213881.2.899

Banga A, Hanssen E, Schreuder R, Neijt A (2012) How subtle differences in orthography influence conceptual interpretation. Written Language and Literacy 15(3):185–208

Banga A, Hanssen E, Neijt A, Schreuder R (2013a) Preference for linking element-en-in Dutch noun-noun compounds: native speakers and second language learners of Dutch. Morphology 23(1):33–56, DOI 10.1007/s11525-013-9211-y

Banga A, Hanssen E, Schreuder R, Neijt A (2013b) Two languages, two sets of interpretations: language-specific influences of morphological form on Dutch and English speakers' interpretation of compounds. Cognitive Linguistics 24(2):195–220, DOI 10.1515/cog-2013-0007

Baroni M, Bernardini S, Ferraresi A, Zanchetta E (2009) The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43(3):209–226, DOI 10.1007/s10579-009-9081-4

Biber D (1993) Representativeness in corpus design. Literary and Linguistic Computing 8(4):243–257, DOI 10.1007/978-0-585-35958-8_20

Bochner H (1984) Inflection within derivation. The Linguistic Review 3:411–421

Bortz J (2005) Statistik für Human- und Sozialwissenschaftler, 6th edn. Springer, Heidelberg, DOI 10.1007/bf01513475

Bresnan J (2007) Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In: Featherston S, Sternefeld W (eds) Roots: Linguistics in Search of Its Evidential Base, De Gruyter Mouton, Berlin/New York, Studies in Generative Grammar, pp 77–96, DOI 10.1515/cllt.2011.011

Bresnan J, Cueni A, Nikitina T, Baayen RH (2007) Predicting the dative alternation. In: Bouma G, Krämer I, Zwarts J (eds) Cognitive foundations of interpretation, Koninklijke Nederlandse Akadmie van Wetenschappen, Amsterdam, pp 69–94

Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science 1(3):140,216, DOI 10.1098/rsos.140216

Cumming G (2014) The new statistics: why and how. Psychological Science 25(1):7–29, DOI 10.1177/0956797613504966

Divjak D, Dąbrowska E, Arppe A (2016) Machine meets man: evaluating the psychological reality of corpus-based probabilistic models. Cognitive Linguistics 27(1):1–33, DOI 10.1515/cog-2015-0101

Dressler WU, Libben G, Stark J, Pons C, Jarema G (2001) The processing of interfixed German compounds. Yearbook of Morphology 1999:185–220, DOI 10.1007/978-94-017-3722-7_8

Dudenredaktion (ed) (2006) Duden: Die deutsche Rechtschreibung, vol 1, 24th edn. Bibliographisches Institut & F. A. Brockhaus, Mannheim

Evert S, Lüdeling A (2001) Morphological productivity: is automatic preprocessing sufficient? In: Rayson P, Wilson A, McEnery T, Hardie A, Khoja S (eds) Proceedings of the Corpus Linguistics 2001 Conference, Lancaster University, Lancaster, pp 167–175, DOI 10.1111/lang.12231

Fehringer C (2009) Wie wär's mit einem richtigen Mädelsabend? Plural -s within compounds in colloquial northern German. Journal of Germanic Linguistics 21(2):149–165

Fisher RA (1926) The arrangement of field experiment. Journal of the Ministry of Agriculture of Great Britain 33:503–513, DOI 10.1007/978-1-4612-4380-9_8

Ford M, Bresnan J (2013) Using convergent evidence from psycholinguistics and usage. In: Krug M, Schlüter J (eds) Research Methods in Language Variation and Change, Cambridge University Press, Cambridge, MA, pp 295–312, DOI 10.1017/cbo9780511792519.020

Fuhrhop N (1996) Fugenelemente. In: Lang E (ed) Deutsch-typologisch, De Gruyter, Berlin, pp 525–549

Fuhrhop N, Kürschner S (2015) Linking elements in Germanic. In: Mueller PO, Ohnheiser I, Olsen S, Rainer F (eds) Word-formation: an international handbook of the langauges of Europe, 1, De Gruyter, Berlin, pp 568–582

Gaeta L, Schlücker B (eds) (2012) Das Deutsche als kompositionsfreudige Sprache: strukturelle Eigenschaften und systembezogene Aspekte. De Gruyter, Berlin

Gallmann P (1998) Fugenmorpheme als Nicht-Kasus-Suffixe. In: Variation und Stabilität in der Wortstruktur: Untersuchungen zu Entwicklung, Erwerb und Varietäten des Deutschen und anderer Sprachen, Olms, Hildesheim, pp 177–190

Gelman A, Geurts HM (2017) The statistical crisis in science: how is it relevant to clinical neuropsychology? The Clinical Neuropsychologist 31(6–7):1000–1014, DOI 10.1080/13854046.2016.1277557

Gelman A, Loken E (2014) The statistical crisis in science. The American Scientist 102(6):460–465, DOI 10.1511/2014.111.460

Gigerenzer G (2004) Mindless statistics. The Journal of Socio-Economics 33:587–606, DOI 10.1016/j.socec.2004.09.033

Goodman S (2008) A dirty dozen: twelve p-value misconceptions. Seminars in Hematology pp 135–140, DOI 10.1053/j.seminhematol.2008.04.003

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31:337–350, DOI 10.1007/s10654-016-0149-3

Haspelmath M (2010) Framework-free grammatical theory. In: Heine B, Narrog H (eds) The Oxford handbook of grammatical analysis, Oxford University Press, Oxford, pp 341–465

Hay JB, Baayen RH (2005) Shifting paradigms: gradient structure in morphology. Trends in Cognitive Sciences 9(7):342–348, DOI 10.1016/j.tics.2005.04.002

Hintze JL, Nelson RD (1998) Violin plots: A box plot-density trace synergism. The American Statistician 52(2):181–184, DOI 10.2307/2685478

Hope ACA (1968) A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society Series B (Methodological) 30(3):582–598

Hunston S (2008) Collection strategies and design decisions. In: Lüdeling A, Kytö M (eds) Corpus linguistics. An international handbook, Walter de Gruyter, Berlin, pp 154–168, DOI 10.1017/cbo9781139524773.003

Kilgarriff A, Baisa V, Bušta J, Jakubíček M, Kovář V, Michelfeit J, Rychlý P, Suchomel V (2014) The Sketch Engine: ten years on. Lexicography pp 1–30, DOI 10.1007/s40607-014-0009-9

Kirchner R, Nicoladis E (2009) A level playing-field: perceptibility and inflection in English compounds. Canadian Journal of Linguistics 54(1):91–116

Koester D, Gunter TC, Wagner S, Friederici AD (2004) Morphosyntax, prosody, and linking elements: the auditory processing of German nominal compounds. Journal of Cognitive Neuroscience 16(9):1647–1668

Köpcke KM (1993) Schemata bei der Pluralbildung im Deutschen: Versuch einer kognitiven Morphologie. Narr, Tübingen

Köpcke KM (1995) Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache – Ein Beispiel für die Leistungsfähigkeit der Prototypentheorie. Zeitschrift für Sprachwissenschaft 14(2):159–180, DOI 10.1515/zfsw.1995.14.2.159

Krott A, Schreuder R, Baayen RH, Dressler WU (2007) Analogical effects on linking elements in German compound words. Language and Cognitive Processes 22(1):25–57

Kupietz M, Belica C, Keibel H, Witt A (2010) The German reference corpus DeReKo: A primordial sample for linguistic research. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10), European Language Resources Association (ELRA), Valletta, Malta, pp 1848–1854, DOI 10.1075/lis.28.08bel

Kürschner S (2005) Verfugung-s-nutzung kontrastiv: zur Funktion der Fugenelemente im Deutschen und Dänischen. Tijdskrift voor Skandinavistiek 26:101–125

Leech G (2007) New resources or just better old ones? The Holy Grail of representativeness. In: Hundt M, Nesselhauf N, Biewer C (eds) Corpus linguistics and the web, Rodopi, Amsterdam and New York, pp 133–149, DOI 10.1163/9789401203791_009

Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? Journal of the American Statistics Association 88:1242–1249, DOI 10.1007/978-1-4614-1412-4_19

Lehmann EL (2011) Fisher, Neyman, and the creation of classical statistics. Springer, New York, NY, DOI 10.1007/978-1-4419-9500-1

Libben G, Jarema G, Dressler WU, Stark J, Pons C (2002) Triangulating the effects of interfixation in the processing of German compounds. Folia Linguistica 36(1-2):23–43

Maxwell SE, Delaney HD (2004) Designing experiments and analyzing data: a model comparison perspective. Taylor & Francis, Mahwa, New Jersey, London

Mayo DG (1996) Error and the growth of experimental knowledge. University of Chicago Press, Chicago, DOI 10.7208/chicago/9780226511993.001.0001

McEnery T, Xiao R, Tono Y (2006) Corpus-based language studies. An advanced resource book. Routledge, London and New York, DOI 10.1111/lang.12225
Mohanan KP (1986) The theory of lexical phonology. Reidel, Dordrecht
Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA (2017) A manifesto for reproducible science. Nature Human Behaviour p 0021 EP, DOI 10.1038/s41562-016-0021
Neef M (2015) The status of so-called linking elements in German: arguments in favour of a non-functional analysis. Word Structure 8(1):29–52
Neef M, Borgwaldt SR (2012) Fugenelemente in neugebildeten Nominalkomposita. In: Gaeta and Schlücker (2012), pp 1–25
Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society A 236(767):333–379, DOI 10.1098/rsta.1937.0005
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A 231:694–706, DOI 10.1007/978-1-4612-0919-5_6
Nübling D, Szczepaniak R (2013) Linking elements in German: origin, change, functionalization. Morphology 23(1):67–89
Peirce JW (2007) Psychopy – Psychophysics software in Python. Journal of Neuroscience Methods 162(1–2):8–13, DOI 10.1016/j.jneumeth.2006.11.017
Perezgonzalez JD (2014) A reconceptualization of significance testing. Theory & Psychology 24(6):852–859, DOI 10.1177/0959354314546157
Perezgonzalez JD (2015a) Confidence intervals and tests are two sides of the same research question. Frontiers in Psychology DOI 10.3389/fpsyg.2015.00034
Perezgonzalez JD (2015b) Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology 6(223):1–11, DOI 10.3389/fpsyg.2015.00223
Perkuhn R, Keibel H, Kupietz M (2012) Korpuslinguistik. Fink, Paderborn
Piantados ST (2014) Zipf's word frequency law in natural language: a critical review and future directions. Psychonomic Bulletin & Review 21(5), DOI 10.3758/s13423-014-0585-6
Pinker S (1999) Words and rules: the ingredients of language. Basic Books, New York
Pollard C (1996) The nature of constraint-based gammar. In: Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation (PACLING'96), DOI 10.1093/ojlr/rww025
R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, DOI 10.1111/j.1467-8624.2009.01290.x
Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape (with discussion). Applied Statistics 54:507–554
Schäfer R (2016) Prototype-driven alternations: the case of German weak nouns. Corpus Linguistics and Linguistic Theory ahead of print, DOI 10.1515/cllt-2015-0051
Schäfer R, Bildhauer F (2012) Building large corpora from the web using a new efficient tool chain. In: Calzolari N, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Odijk J, Piperidis S (eds) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12), ELRA, Istanbul, pp 486–493
Schäfer R, Bildhauer F (in prep.) The COW16 web corpora. In preparation (draft available on request from the authors)
Scherer C (2012) Vom Reisezentrum zum Reise Zentrum – Variation in der Schreibung von N+N-Komposita. In: Gaeta and Schlücker (2012), pp 57–81
Schlücker B (2012) Die deutsche Kompositionsfreudigkeit: Übersicht und Einführung. In: Gaeta and Schlücker (2012), pp 1–25
Schmid H, Fitschen A, Heid U (2004) SMOR: A German computational morphology covering derivation, composition, and inflection. In: Lino MT, Xavier MF, Ferreira F, Costa R, Silva R (eds) Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 4), Universidade Nova de Lisboa, Lisbon, pp 1263–1266, DOI 10.1075/z.139.28hei
Schreuder R, Neijt A, van der Weide F, Baayen RH (1998) Regular plurals in Dutch compounds: linking graphemes or morphemes? Language and Cognitive Processes 13(5):551–573
Senn SJ (2011) You may believe you are a Bayesian but you are probably wrong. Rationality, Markets, and Morals 2:48–66, DOI 10.1002/ss.20167
Šidák ZK (1967) Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633, DOI :10.1080/01621459.1967.1048293
Siegel D (1979) Topics in English morphology. Garland, New York
Stefanowitsch A, Flach S (2016) A corpus-based perspective on entrenchment. In: Schmid HJ (ed) Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning, De Gruyter, Berlin, pp 101–128, DOI 10.1037/15969-006

Szczepaniak R (2016) Is the development of linking elements in German a case of exaptation? In: Norde M, de Welde FV (eds) Exaptation and language change, Benjamins, Amsterdam, pp 317–340, DOI 10.1075/cilt.336.11szc

Trafimow D, Marks M (2015) Editorial. Basic and Applied Social Psychology 37(1):1–2, DOI 10.1080/01973533.2015.1012991

Verhoeven E, Temme A (2017, to appear) Word order acceptability and word order choice. In: Featherston S, Hörnig R, Steinberg R, Umbreit B, Wallis J (eds) Linguistic Evidence 2016 Online Proceedings, Univeristät Tübingen, Tübingen, DOI 10.1515/ling-2016-0018

Wegener H (2003) Entwicklung und Funktion der Fugenelemente im Deutschen. Oder: Warum wir keine *Autosbahn haben. Linguistische Berichte 196:425–456

Wellmann H, Reindl N, Fahrmeier A (1974) Zur morphologischen Regelung der Substantivkomposition im heutigen Deutsch. Zeitschrift für deutsche Philologie 93:358–378

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, Berlin etc., DOI 10.1007/978-0-387-87458-6