2

I'M LEAKING OIL AND LOOKING FOR A GARAGE: TESTING CONDITIONS ON MEANING TRANSFER

Sam Featherston, Klaus von Heusinger and Hanna Weiland

ABSTRACT

In this chapter we report experiments aiming to verify the conditions under which meaning shift can occur. We address claims by Nunberg (1995, 2004) that a salient "functional relationship" and "noteworthiness" between the primary meaning and the extended meaning are prerequisites of shifting and that such meaning shift should preferentially be analyzed as occurring in such a way as to preserve established reference assignment. As an example case of this we looked at German deverbal nominalizations in *-ung* which have both Event and Result readings, building on the work of Brandtner and von Heusinger (2010). The results reveal that the effects predicted by Nunberg are psychologically real, but that they are not specific to the "reading shift" environment; rather they are a background effect of coherence. In this paper we particularly focus on the process of

Experiments at the Interfaces
Syntax and Semantics, Volume 37

ISSN: 0092-4563/doi:10.1108/S0092-4563(2011)0000037006

testing these claims: such extensive control of the lexical materials requires that the generalizability of the results be thrown into question. We discuss the implications of these facts for empirical verification in questions of interpretation.

1. INTRODUCTION

It is a central characteristic of language use that utterances can be interpreted not literally but in shifted senses. However, in spite of much work by researchers in both linguistics and literature, the pragmatic conditions which license such meaning shifts are only poorly understood. Nor do we have clear criteria to determine what part of an utterance receives a shifted interpretation. For example, the utterance in (1) must be understood in a transferred sense: a person cannot be parked anywhere; only vehicles can be parked. So what is the speaker of this surprisingly natural statement saying?

(1) I am parked on a pedestrian crossing.
  a. My car is parked on a pedestrian crossing.
  b. I am the driver of a car which is parked on a pedestrian crossing.

We may identify two ready possibilities: (1a) and (1b). In the first case, it is the reference of the subject which is given a shifted interpretation, so that *I* is understood as *my car*. In the second, the predicate is enriched to be a feasible property of a person. On the face of it, either of these might be the intended communicative content of (1) (Nunberg, 1995, 2004; Weiland, Featherston, & von Heusinger, 2010). Nunberg (1995) discusses some tests that show that reading (1b) is the most probable one (see below).[1]

This chapter has two aims. First, some studies which address the claims about such "predicate transfers" made by Nunberg (1995, 2004). Second, we discuss the particular methodological challenges in constructing a valid experimental investigation of this sort. We shall amplify these two aims in turn.

Nunberg proposes two hypotheses about the conditions and domains of meaning shift in pragmatic interpretation. The first concerns the part of an utterance that receives a shifted interpretation. For example, in *I am parked on a pedestrian crossing* it might be assumed that it is the subject pronoun which receives a shifted interpretation, since the statement as a whole concerns the location of the car, not of the owner. Nunberg argues on the other hand that we may usefully think of the predicate as having the transferred interpretation,

[1]This is an adaptation of Nunberg (1995, 2004) primary example *I am parked out back*, which is unfortunately not well understood internationally.

since the location of the car is a noteworthy piece of information about the person who drives it (see also Copestake & Briscoe, 1995).

This leads us on to the second hypothesis, which is a condition on such pragmatic meaning shifts. Nunberg suggests that these are only possible when there are particular sorts of relations between the literal and transferred bearers of the properties, such as between a driver and a car. We may note that many other apparently not dissimilar examples seem much less natural or even impossible, which demonstrates the existence of quite strict conditions on such transfer. Even close parallels to drivers and cars, such as cyclists and bicycles, seem to allow such transfer much more restrictedly (2).

(2) ?I am locked to the railings outside. (where "I" refers to "my bicycle")

The research questions on meaning transfer that we address are thus the circumstances which license such shifts of meaning, and the specification of which part of an expression receives a shifted interpretation. To do this we focus on a specific instance of the phenomenon, namely the polysemy of nominalizations in German.

German deverbal nominalizations with *-ung* can have different readings depending on the context they appear in. A problem appears if separate parts of the surrounding context trigger different readings, as in (3a) and (3b). In (3a) *langwierig* "time-consuming" indicates an Event reading for *Übersetzung* "translation," while the predicate *lag auf dem Tisch* "lay on the table" suggests a Result reading. In (3b) the adjective *abblätternd* "flaking" triggers a Result reading for *Bemalung* "painting," while the predicate *dauerte lange* "took a long time" indicates an Event. We shall refer to the parts of the context which trigger specific readings of the head noun as "indicators." In the examples here, there is always one indicator which is a premodifier and precedes the head noun, and one which is a predicate and follows it (see Brandtner & von Heusinger (2010) and Brandtner (2011) for discussion).

(3) a. Die [langwierige]*Ev* Übersetzung [lag auf dem Tisch]*Res*.
"The time-consuming translation lay on the table."
b. Die [abblätternde]*Res* Bemalung [dauerte lange]*Ev*.
"The flaking painting took a long time."

Such examples provide an interesting test case for the conditions and domains of meaning transfer suggested by Nunberg (1995, 2004), as the manipulation of the two indicators allows us full control over the reading of the nominalization, which is a precondition for investigating meaning shift.

This therefore is the linguistic content of our studies. We shall report and discuss the findings of our studies in this chapter, but we cover them in greater detail in Weiland et al. (2010). The central concern of this chapter is the construction process of the material for these experiments, because it provides

useful lessons in the data basis which is required to support a hypothesis of this type. In particular, it is relevant to the testing of hypotheses which make reference to either lexical sets (as here the German nominalizations in *-ung*) or to aspects of meaning. Our finding is that there are inherent difficulties in experimentally testing such hypotheses, which may require some rethinking about how the value of such theories is assessed, if we accept that verifiability is a precondition of meaningful theory construction.

There are three main methodological points which we wish to raise here. The first is the difference between "preferred interpretation" and "forced interpretation," which must, it turns out, be strictly distinguished. For example in (4) the German phrase *Geld auf der Bank* ("money in the bank" or "money on the bench") is strongly preferred to have the interpretation where a *Bank* is a financial institution, but it could have the interpretation in which piles of money are upon a park bench, as here. By contrast *Guthaben auf der Bank* ("credit in the bank") in (5) probably has the forced interpretation "financial institution."

(4) Das Geld auf der Bank im Stadtpark wurde von einer alten Dame vergessen.
"The money on the bank/bench in the town park was forgotten by an old lady."

(5) ?Das Guthaben auf der Bank im Stadtpark wurde von einer alten Dame vergessen.
"The credit on the bank/bench in the town park was forgotten by a old lady."

The additional language processing involved in a shift from a preferred interpretation is barely perceptible, of a very different magnitude to the processing load of an incompatible forced interpretation. In the first case, we might only be dealing with a sharpening of a previously not fully specified interpretation; in the second we can be sure that a shift in meaning, not just a narrowing, has taken place. This distinction is important if we wish to differentiate between meaning shift and incremental specification, as here.

Our second topic is the difference in approaches to controlling for irrelevant variables in experimental studies. We shall argue that control in such contexts involves a trade-off of two desirable factors: identity across conditions and optimal naturalness for each condition. In our first study series we adopted what we shall refer to as the *local optimum* method of control, and adjusted the context factors for each condition so as to make them as natural as possible for just that condition, even though they thus varied across conditions. In our second approach we placed higher value on what we here dub the *identity* method, that is, we tried to apply control by keeping factors constant, selecting only those lexical variants which were compatible across conditions.

Our third methodological issue is a reflection on the necessity of hypotheses to be testable. It is widely held that meaningful claims in academic work must

at least in principle be verifiable. However, a range of factors make the hypotheses we address here difficult to verify. The process of selection of exemplars of the *-ung* nominalizations is so stringent that the question arises whether the remaining sample tested are representative of the group as a whole. We therefore have another trade-off, this time between control and generalizability: if the degree of restrictiveness required by a hypothesis exceeds a certain point, the materials that fulfill these stringent criteria can no longer claim to be randomly selected, so that the generalizability of our findings is no longer given. In this situation we face a drastic reduction in the value of our investigations: instead of learning about how facets of language in general function, our work may deliver no more than observations about exactly the items tested, no further generalization being possible. These three findings should be of interest to linguists who are concerned about the data basis of their claims.

The structure of the rest of the chapter is as follows. We first give a little more background information about our study series and set our research questions in context. Next we describe our first study series using the "local optimum" approach to control. We then show why these results were questioned and describe our second study series, which employed the "identity" method of control. We finish with a survey of our findings but focus on the methodological implications.

2. BACKGROUND

2.1 Nunberg (1995, 2004)

In two papers on meaning shift, Nunberg notes that in a "sortal mismatch" we have to consider which part can be adjusted to the other and under what circumstances. Nunberg refers to examples with a mismatch between a subject personal pronoun *I* and a predicate *be parked*, which normally applies just to cars, as in (6a). Instead of shifting the reference of the pronoun to refer to the car, Nunberg argues that the mismatch is solved in this case by the enrichment of the predicate *be parked* so that it can also apply to persons, as in (6b), while the pronoun still refers to the owner.

(6) a. I am parked on a pedestrian crossing.
b. I am [the owner of a car that is] parked on a pedestrian crossing.

Nunberg (1995, 2004) discusses different tests to distinguish whether the subject or the predicate receives a shifted interpretation. We focus here on the coordination test, also known as "copredication" (Brandtner, 2011; Cruse, 2004). He argues that (6a) undergoes predicate transfer and shows a shifted reading as in (6b), since the subject can also be combined with a predicate that only selects human arguments, as in (7a). If we combine the subject with

a predicate that selects an inanimate machine, as in (7b), the sentence becomes infelicitous, which indicates that the subject is not shifted.

(7) a. I am parked on a pedestrian crossing and have been waiting for 15 minutes.
b. *I am parked on a pedestrian crossing and may not start.

The copredication in (8a) also suggests a predicate transfer from the predicate *be leaking oil* to the extended predicate *be the owner of a car that is leaking oil*, rather than a shift from *I* to *my car*:

(8) a. I'm leaking oil and looking for a garage.
b. I [am the owner of a car that is leaking oil] and looking for a garage.

Another of Nunberg's examples of copredication is (9a), where the first predicate *is Jewish* suggests a Person reading for Roth, while the second one indicates a Book reading, since only books can be read.

(9) a. Roth is Jewish and widely read.
b. Roth is Jewish and is [an author of books which are] widely read.

Nunberg suggests that the first indicator fixes the reading for the proper name here, while the second one is adjusted to these requirements, so that *is widely read* becomes an enriched property applying to persons, as in (9b). Accordingly, we have only one reading for the noun in such sentences, thus solving the incompatibility.

This mechanism is not unconstrained; Nunberg (1995, 2004) suggests that there are two conditions for predicate transfer (10) (see also Copestake & Briscoe, 1995).

(10) a. The bearers of the properties must stand in a "salient functional relation."
b. The property contributed by the new enriched version has to be "noteworthy" for the identification or classification of the bearer.

There is, for example, a salient functional relation between drivers and their cars, and it can certainly be noteworthy for a driver that his car is parked in a particular place.

Nunberg argues that the two constraints are separate, referring to examples such as (11a) and (11b). A property of a car, such as being damaged, would not be noteworthy for a dead car owner, Nunberg suggests, arguing that this lack of noteworthiness explains the unnaturalness of (11b).

(11) a. Ringo was hit in the fender by a truck while he was momentarily distracted by a motorcycle.
b. ??Ringo was hit in the fender by a truck two days after he died.

Nunberg also provides another example (12), said by an artist. He suggests that being in a prestigious gallery is a noteworthy property of a person, being in a crate is not.

(12) a. I am in the Whitney Museum.
b. ?I am in the second crate on the right.

It is clear that there are effects in these examples, but the cause is perhaps not as clear as Nunberg assumes. Since several of Nunberg's examples allow alternative accounts, and since Nunberg himself admits that he can offer no more than an intuitive criterion for noteworthiness, we shall therefore not distinguish between salient functional relations and noteworthy relations but refer only to *Relatedness*. We consider Nunberg's examples in more detail in Weiland et al. (2010). Nevertheless, these cases provide clear evidence of contrasts in acceptability which motivate Nunberg's constraints on the phenomenon, and it is this issue that we address. The questions about the location of and conditions on meaning shift remain.

2.2. German Deverbal Nouns in *-ung*

The case of meaning shift that we address here is that of German deverbal nominalizations in *-ung*, following Brandtner and von Heusinger (2010). German verbal stems can fairly productively be extended with an *-ung* suffix to yield a noun (so *absperr-en* "block up" yields *Absperr-ung* "blocking (up)," and *bearbeit-en* "process" or "deal with" yields *Bearbeit-ung* "processing" or "treatment"). Since they are deverbal, it may be assumed that the primary meaning of these nominalizations relates to the process expressed by the verb. But forms in *-ung* have a variety of different readings; so *Übersetzung* ("translation") can refer to the process of translating or to a translated text, *Verwaltung* ("administration") can refer to a process, the people who do it, and the place where they do it.

Ehrich and Rapp (2000) distinguish types of possible readings for these nominalizations. They suggest a first division into the types Eventuality and Result Object, with the former being further divided into Process, Event, and State readings, this last being subdivided once more into Result State and Nonresult State types — see Figure 1. We shall here make only a single distinction: between Event (Ev) and Result (Res). The first combines the subsorts Event and Process from Ehrich & Rapp's Eventuality group, but not State; the second corresponds to their Result Object sort. The reason for this simplification is that any further degree of distinction is very difficult to reliably achieve in experimental materials except in a few prototypical cases.

Ehrich and Rapp (2000) note that the specific reading of an *-ung* nominal in any individual occurrence is triggered by elements of the context, generally the selectional restrictions of modifiers and predicates, which they call "reading indicators." Factors such as duration as in (13a), but also time frame predicates

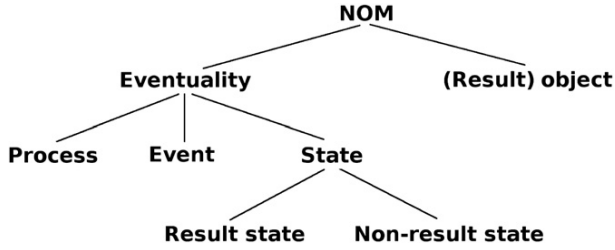

Figure 1: Ehrich and Rapp's (2000) nominalization types.

and dates, for example, require Event readings, while physical change and appearance predicates as in (13b) suggest a Result reading.

(13) a. Die Bemalung [dauerte lange]. (Ev)
"The painting took a long time."
b. Die Bemalung [ist rot-schwarz gestreift]. (Res)
"The painting is striped red and black."

Since the context can specify which reading is intended, the question arises what happens when the contextual clues conflict, as in (3a) and (3b) repeated as (14).

(14) a. Die [langwierige]*Ev* Übersetzung [lag auf dem Tisch]*Res*.
"The time-consuming translation lay on the table."
b. ?Die [abblätternde]*Res* Bemalung [dauerte lange]*Ev*.
"The flaking painting took a long time."

This phenomenon is referred to as "copredication" in Brandtner and von Heusinger (2010) following Asher (2011), and Pustejovsky (1995). Although these authors and Cruse (2004) have recognized and researched the phenomenon with simple nouns, there is no agreement on how to handle it and what follows for a theory of predication. We may distinguish two approaches: the first asks what happens in the on-line processing of such cases, and the second treats it as a question about the semantic representation.

Brandtner and von Heusinger (2010) suggest an account building on the work of Nunberg (1995, 2004), in which it is the predicate which undergoes an adaptation of meaning to match the reading of the noun phrase established by the first indicator, rather than the nominalization itself. This approach also seems plausible in processing terms, since it might be more economical in cognitive resources to adopt a less accessible interpretation of new linguistic input than to adjust our interpretation of previously processed material. Garden paths are an example of this: on reading *The horse raced past the barn fell*, our parser is reluctant to reanalyze the string *the horse raced past the barn* as a noun phrase, even though the final word *fell* tells us that we should. If coercions

on new analyses are cognitively "cheaper" than reanalyses of past input, Nunberg's location of the shift in the predicate would be supported.

However, there are also considerations which support exactly the opposite analysis. It is a common observation that constraints on form or interpretation are applied more strictly immediately that they are met than later, when a degree of "decay" in their activation has occurred. Looking at (15), the reading of the nominalization is determined as an event by the premodifying indicator, and then undergoes "wrap-up" processing at the end of the noun phrase; its meaning is established and digested as a chunk (cf. "sausage machine" Frazier & Fodor, 1978). This could make it easier for the reading of the nominalization to be shifted when the constraints from the next indicator arrive in the input, for the coercion toward an object reading is thus more salient.

(15) The laborious painting ... was on the table.

Our approach in this research was to gather data about the way the mismatch cases are perceived as a first step toward developing a more adequate account of their analysis.

2.3. Brandtner and von Heusinger (2010)

Brandtner and von Heusinger (2010) apply Nunberg's mechanism of predicate transfer to copredication cases in German deverbal nominalizations derived with *-ung*. They explore the possibility of accounting for reference shifts in single utterances by locating them in the verbal predicate instead of in the nominalization. This permits the nominalization to have a single reference, that which is determined by the first indicator. As shown in (17), the enriched version of (16), the nominal has only one fixed reading in this sentence while the predicate part of the context is adjusted to it, so that we have two event predicates applying to the nominal *Übersetzung* "translation."

(16) Die [langwierige]*Ev* Übersetzung [liegt auf demTisch]*Res*.
"The time-consuming translation is lying on the table."

(17) Die [langwierige]*Ev* Übersetzung [*hat ein Resultat, das* auf dem Tisch liegt]*Ev*.
"The time-consuming translation had a result that is lying on the table."

Brandtner and von Heusinger note that one can assume there always to be a close relation between events and their results, so that the multiple readings of deverbal nominals are automatically in a functional relation. Effectively therefore the question becomes rather whether there is perceptible Relatedness between the properties added by the indicators. We see the effect in (18) and its absence in (19).

(18) Die [abblätternde]*Res* Bemalung [wurde schlampig durchgeführt]*Ev*.
"The flaking painting was carried out sloppily."

(19) ?Die [abblätternde]*Res* Bemalung [dauerte lange]*Ev.*
"The flaking painting took a long time."

Both examples contain a mismatch between the first indicator *abblätternd* "flaking" triggering a Result reading and the second triggering an Event reading. Brandtner & von Heusinger point out that in (18) there is a connection between the fact that the paint is flaking and how the painting was done, since the second can account for the first. There is no such plausible connection in (19), however, since more time taken does not usually give lower quality results. Felicitous copredication thus seems to depend not only on the semantic content of the nominalization, but also requires there to be a plausible Relatedness relation between the indicators (10).

In this chapter we attempt to verify these assumptions by testing them across lexical sets using judgement studies. We first test whether the assumptions and predictions in the literature so far can be confirmed in experimental data and then consider which accounts are supported. We use acceptability judgement studies to test first, what influence the different combinations of reading indicators have on the acceptability of mismatched structures, and second, what the role of Relatedness is.

2.4. Aims and Predictions

The overarching aim of these studies was to address three issues from Nunberg (1995):

i. Pragmatic meaning shift: Does Relatedness (functional relation, noteworthiness) support the coercion process of predicate transfer?
ii. Meaning representation: Are the meanings underspecified, or fully spelled out, but "densely metonymous"?
iii. Is there evidence for a directional derivation from Event to Result?

We must approach the summits of these linguistic aims with a long approach march through the foothills of testable predictions. We first note that for any hypothesis such as this one concerning a lexical group, we must be able to find a pattern of behaviour amongst them. This is not trivial, as the shared derivational pattern does not necessarily determine synchronic behaviour. We should next wish to verify the simple prediction that examples with two indicators for the same reading will be judged better than examples with nonmatching indicators such as (18) and (19). Relating to derivational direction, the prediction would be that among the sentences with nonmatching indicators, those with the Event reading as the first indicator and the Result reading as the second indicator will be more acceptable that those with the inverse ordering. The basis of this prediction is that the event before result sequence corresponds to the inherent ordering of events and results of events.

Another factor which might favour such an effect is the direction of derivation. The *-ung* nominalizations are deverbal, which implies that the Event readings are logically prior to the Result readings (Brandtner & von Heusinger (2010), but see Cruse (2004) for a critique).

Nunberg's conditions in (10) predict that copredication will only be felicitous if there is Relatedness across the parts of the sentence: between the first indicator and nominalization on the one hand, and between these two taken together and the second indicator on the other. It is therefore necessary to test whether copredication examples with internal Related parts are judged more acceptable than equivalent examples without. In the following we report our two series of studies designed to investigate these questions.

3. EXPERIMENT SERIES 1: LOCAL OPTIMAL CONTROL

3.1. Methodological Considerations 1

Creating the materials for these experiments required extensive pretesting of the linguistic materials, which we partially report here. It also required us to prioritize what factors most needed to be controlled for. We constructed the materials so that all examples are maximally plausible, so as to prevent differences in plausibility causing distortion in the results. This is necessary in experimental designs in which a lexical factor is a condition, as here. This point may require some illustration: if we are testing a structural difference — say between goal arguments as prepositional phrases or as shifted datives in English — then it is fundamentally the case that just the structure counts, and we should find the same effect in any lexical material, as long as it is matched for length, frequency, plausibility, and phonotactics. In (20) we can put in any related set of noun phrases and the effects should be consistent. Testing structural questions is thus fairly straight forward.

(20) The teacher/sergeant/bishop gave the student/soldier/curate a pencil/rifle/bible.

This changes radically if we test across a lexical set, for example across ditransitive verbs. While just about anyone can *give* or *send* something to someone else, it is much more restricted who can *throw*, *push*, *sell*, *drag*, *take*, *fax*, *toss*, *flip*, *slap*, *kick*, *poke*, *fling*, *blast*, *carry*, *pull*, *lift*, *lower*, *haul* a given thing to another (partial verb list from Bresnan & Nikitina (2009)). Anything *hauled* must be heavy, anything *lowered* must be fairly heavy and it must previously have been in higher up, and anything *flung* is being treated with contempt. This massively reduces our choice of lexis (21).

(21) a. The teacher ??lowered/??hauled/flung/flipped/*blasted the student a pen.
b. The sergeant lowered/?hauled/flung/?flipped/*blasted the soldier a rifle.
c. The bishop ??lowered/*hauled/??flung/*flipped/*blasted the curate a bible.

If we wish to test across the set of verbs which allow dative shift in order to generalize or, worse still, in order to identify the extent to which the individual verbs permit dative shift, the task of materials construction is transformed. We must adopt a different model of control and the primary aim must be to find any example sentence with a similar structure and length to the others in which a person can plausibly *lower* (for instance) anything to anyone else, perhaps (22). This context won't work for *fling* or *flip* or *blast*, though it might for *haul* and *lift*.

(22) The water engineer lowered the technician the instrument.

In practice we have to write more or less a new sentence for each verb, so that control takes on a new form. We call this *local optimum* control, as the task becomes one of making all example sentences equally plausible, not by making them identical — which is impossible — but by finding the best possible context for each of them.

We are forced into this alternative approach to control whenever we wish to test a lexical variable as one of our experimental conditions. If a condition constrains the lexis of the items, the clean distinction of conditions and items disappears, so there can be no separate "by items" analysis, and strict control of items becomes impossible: it is fundamentally more marked to *flip*, *haul*, or *fling* something than it is to *give*, *hand*, or *send* it. One of the aims of this chapter is to consider the implications of these facts for materials construction in experimental studies like this.

3.2. Pretests

We carried out several preparatory studies in order to develop the materials. Space does not permit full details here, but we shall sketch just two of them. The aim of the first preparatory study was to identify suitable nominalizations in *-ung* for our experiments. Since the aim of the research is to identify the circumstances that permit or favour a reinterpretation of these nouns from one reading to another, those we test must have equal background acceptability on both readings, and in many cases one of the two readings is strongly lexicalized. For example *Werbung* "advertising" in German is generally interpreted as a result, as the outcome of the activity. On the other hand, *Lesung* "reading" has a fairly robustly lexicalized event reading. Neither of these would therefore be suitable for our purposes.

We gathered acceptability judgements of 40 candidate nominalizations together with indicators of either Event or Result readings that either precede or follow the head noun. We thus used a 2 × 2 design with two factors: indicator reading (Result, Event) and indicator position (premodifying adjective in NP, VP following) (Table 1). We also tested control sentences containing *-ung* nominalizations with a strongly lexicalized reading, such as *Wohnung* ("apartment").

The 50 native speakers of German who took part were recruited by e-mail at the University of Stuttgart. The instructions told them to read the sentences carefully and judge them spontaneously on a four-point labelled scale, from "sounds very good" to "sounds very bad." The results allowed us to form a pool of 22 nominalizations which were judged first, nearly as good as the "good" control conditions, and second, roughly equally acceptable on the two readings.

The aim of the second preparatory study was to test whether the experimenters' opinion of what constitutes a Relatedness relation would be confirmed by a wider range of informants. We tested a sample of the indicators that we had developed from the suggestions in the literature (Brandtner & von Heusinger, 2010; Ehrich & Rapp, 2000) and applied them to a sample of the nominalizations confirmed to be usable in the first preparatory study. The task was to choose one of two possible sentence continuations (Table 2).

Table 1: Sample experimental items in pretest on balanced nominalizations.

Conditions and example stimuli	
NP Res die abblätternde Bemalung "the flaking painting"	*VP Res* Die Bemalung besteht aus alter Ölfarbe. "The painting consists of old oil paint."
NP Ev die gemeinsame Bemalung "the collective painting"	*VP Ev* Die Bemalung dauerte lange. "The painting took a long time."

Table 2: Sample experimental items in pretest on Relatedness.

Indicators	Example stimulus	Ending type	Example ending.
NP Res VP Ev	Die verschwundene Erzählung "The lost narration	Related	. . . wurde nicht beendet . . . was not finished"
		Unrelated	. . . fand gestern statt ". . . took place yesterday"
NP Ev VP Res	Die gestrige Erzählung "Yesterday's narration	Related	. . . wurde heute illustriert . . . was illustrated today"
		Unrelated	. . . liegt auf dem Tisch ". . . is lying on the table"

Each of the 10 nominalizations was tested with just one premodifying indicator each for the Event and Result readings (*NP Ev*, *NP Res*), which provided 20 sentence beginnings. For each of these we constructed two sentence endings, both of which contained a mismatching VP indicator (so *NP Ev*→*VP Res*, and *NP Res*→*NP Ev*) but only one of which had a Relatedness relation to the first part. Participants were instructed to choose the continuation of the sentence which was more meaningful to them. The 20 experimental items were mixed among 12 control items which contained words ending with *-ung*, but which have only one, strongly lexicalized meaning. These provided points of comparison for acceptable and unacceptable examples. The results revealed that the perception of "relatedness" of the experiment designers was generally, but not exclusively, shared by the experiment participants. This information was used to improve the materials.

3.3. Main Study on Meaning Shift 1

Our preparatory studies allowed us to produce the materials for a controlled experimental study to test the effects of matching versus nonmatching indicators, before and after the nominalization. More formally, we distinguish two parameters, that of the reading of the indicator (Result, Event) and of its position (in the NP, in the VP). We also wished to test the effect of Relatedness and therefore distinguish one more parameter (Relatedness, no Relatedness). We only test this for those combinations of indicators with a mismatch of indicator readings, since Nunberg makes no predictions for examples with matching indicators. This experimental design has thus three factors: Indicator_1 (Result, Event), Indicator_2 (Result, Event), and Relatedness (Related, non-Related).

For this experiment the best 18 nominalizations from preparatory study 1 were used in sentences which were constructed for each of the six experimental conditions (Table 3). The resulting materials were distributed over six versions of the experiment so that the participants in the experiment see each of the six conditions three times but each nominalization only once. The 18 experimental sentences were mixed with 12 filler sentences using strongly lexicalized words with the ending *-ung*, as in the second preparatory study above.

This experimental questionnaire was administered on-line using the WebExp2 experimental package (Keller, Gunasekharan, Mayo, & Corley, 2009). Participants were students at the University of Stuttgart and were recruited by e-mail, giving the URL of the start page of the experiment. This page introduced the study and explained the nature of the task. The second page contained a java applet within which the experiment screens appear. Participants were asked to provide their name, age, dialect, occupation, sex,

Table 3: Sample experimental items per condition in main study 1.

Ind_1	Ind_2	Related	Example stimulus
Result	Result	∅	Die abblätternde Bemalung besteht aus alter Ölfarbe. "The flaking painting consists of old oil paint."
Result	Event	Related	Die abblätternde Bemalung wurde schlampig ausgeführt. "The flaking painting was sloppily carried out."
Result	Result	Unrelated	Die abblätternde Bemalung dauerte lange. "The flaking painting took a long time."
Event	Event	∅	Die gemeinsame Bemalung dauerte lange. "The collective painting took a long time."
Event	Result	Related	Die gemeinsame Bemalung besteht aus Fingerfarbe. "The collective painting consists of finger paints."
Event	Result	Unrelated	Die gemeinsame Bemalung besteht aus alter Ölfarbe. "The collective painting consists of old oil paint."

handedness, and e-mail address. The experiment commenced with a practice phase.

During the experiment, sentences were displayed on the screen in random order, together with a representation of a four-point judgement scale. Participants were instructed to choose a value on the scale for each sentence. A total of 310 people participated in this study, of whom 11 were not native speakers of German and 12 had more than 25% incorrectly answered filler sentences. Of the remainder we selected the first 48 as our result set. For analysis we assigned the numerical values 1–4 to the labelled points "sounds very good," "sounds good," "sounds bad," and "sounds very bad" on the judgement scale. We present the results graphically in Figure 2.

3.4. Discussion of Results

At first sight, these results generally confirm the assumptions in the literature. First of all, the lexical set of nominalizations does seem to be responding consistently, so that we can meaningfully talk about a group result. This would be predicted, since the deverbal derivation is still active and transparent to speakers, but it need not be the case. Second, we notice that the examples with matching indicators are judged better (*Res Res* and *Ev Ev* > *Ev Res* and *Res Ev*), which would be unpredicted if we assumed that the nominalizations are initially given an underspecified interpretation. Third, the effect of Relatedness is well supported, since *Rel+* conditions are consistently better than *Rel−* conditions. Fourth, these results would tend to support the direction of derivation effect, since *Ev Res* conditions are better than *Ev Res* conditions.

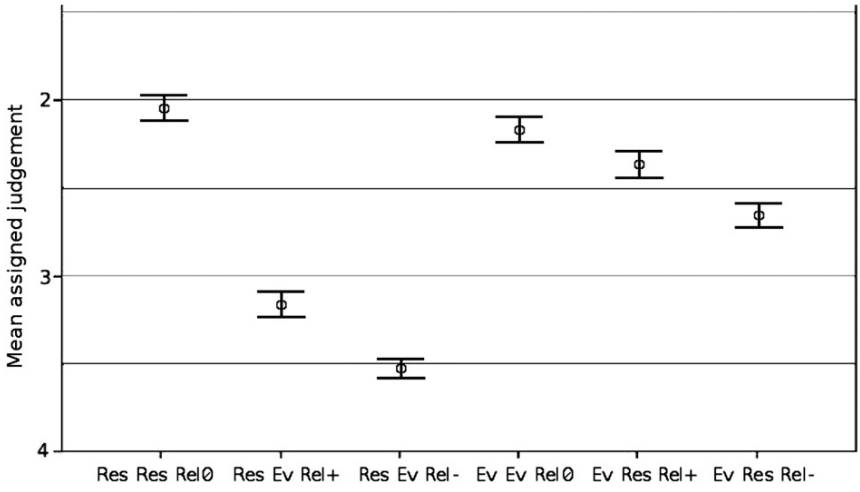

Figure 2: Experiment 1: Means and 95% confidence intervals by condition.

There are however also some puzzles; above all, the fact that the drop in acceptability from the conditions with matching indicators (*Ev Ev*, *Res Res*) to the conditions with nonmatching indicators (*Res Ev*, *Ev Res*) is not consistent. The drop from *Res Res* to *Res Ev* is much larger than from *Ev Ev* to *Ev Res*. This is unexpected, since the effect of a reanalysis would be predicted to be fairly constant. One possible account would be that participants preferred the *Ev Res* order of indicators because this corresponds to the chronological order of Event before Result.

In order to identify the cause of the effects in the results, we looked at the item-specific results and carefully reassessed the materials. This process made us aware that some of the reading indicators were less equivocal than we might have wished. Indeed it turned out that under contextual pressure almost all indicators can tolerate a reading other than their most accessible reading, but in these results it was particularly clear that our premodifying Event indicators could also be quite readily given a Result interpretation, thus annulling the indicator mismatch.

An example of this is *gemeinsam* ("collective," "common," "joint"), as in *die gemeinsame Bemalung der Wand* ("the collective painting of the wall"), which awakens the image of a group of people painting together, an event therefore. On further consideration it became clear that terms such as *gemeinsam* can also apply to things which are clearly not events, such as *unsere gemeinsame Veröffentlichung* ("our collective publication") or *unsere gemeinsame Wohnung* ("our joint home"). Another example is *heutig* ("today's"), as in *die heutige Darstellung* ("today's presentation"). Although, as Ehrich and Rapp (2000) correctly note, an association with a time can function as an indicator of

Eventness; reflection shows that nonevent readings are possible, for example, *die heutige Zeitung* ("today's paper") is as natural in German as it is in English. Another instance is *angefangen* ("started," "commenced"), which most naturally applies to processes, hence Events. But a cake or a book or a statue can all be *angefangen*, meaning that one has started to eat or read or sculpt them.

These facts affected the pattern of our results. What we are measuring in indicator mismatch examples is the cost in perceived well-formedness of forced reinterpretation. This could be the reinterpretation of either (a) the NP consisting of the first indicator and the head noun, or else (b) the second indicator VP. Since the interpretative system has no way in advance of knowing which will first yield a reading compatible with the other, we must imagine two strands of linguistic processing at the same time, one searching for a reading of indicator 1 compatible with indicator 2, and the other searching for a reading of indicator 2 compatible with indicator 1. Taking these considerations into account, it is evident that our measure is a complex one, and very dependent on the individual reading preferences of the lexical items involved. Since reflection has shown that many of our premodifying Event indicators (*heutig*, *gemeinsam*, ...) have dispreferred but still relatively accessible Result readings, this would account for the pattern of results showing *Ev Res* examples as much better than *Res Ev*. The results do not therefore demonstrate the effects that we had first assumed.

3.5. Methodological Considerations 2

This motivated a rather different approach to materials creation in the second experiment series. Rather than employ indicators which yielded a preferred reading, we decided to select only those which forced just one reading. This excludes very many indicators; for example, more or less only indications of duration seem to force an Event reading. The use of unambiguous indicators should provide us with an absolute measure of the difficulty of analysis of noncompatible readings, which will permit us to identify whether in fact there is a preference for *Res Ev* over *Ev Res*, or whether the effect observed in the main experiment was merely a measure of the ease of reanalyzing the specific lexical items tested, as we suspect.

An additional aim of our second experimental series was to produce evidence which can be reasonably generalized. This is scarcely possible on the basis of the experimental materials so far, because the lexis varies too greatly between conditions. We therefore also adopted a different approach to control in these materials. In the first experiment we employed the *local optimum method*, making the lexical material as good as possible in each condition. In this follow-on experiment we employed the *identity method*, in which all conditions are judged with exactly the same lexical material, as far as this is possible. This second method is common and preferred in studies of language structure, but the local optimum method is more often applied in tests of lexical sets. Since

implausibility has such a very strong effect upon judgements, it is important to avoid it as far as possible, even at the cost of a degree of lexical difference between conditions. It was however evident that in this case we needed to reduce the effects of lexical variation on our results.

Another factor which motivated further work was the recognition that it was not sufficient to demonstrate the effect of Relatedness in cases of meaning shift which Nunberg predicted, it was also necessary to show that it was differentially present in such cases. Internal relatedness raises the coherence and thus the acceptability of any example. For this reason experimenters standardly use examples with apparent coherence to avoid shock effects. Examples (23) and (24) are both coherent and will be judged better than if any one element is extracted from one and inserted into the other.

(23) The bishop told the curate to read the prayer book/#clean the machine gun.

(24) The sergeant told the soldier to clean the machine gun/#read the prayer book.

We must therefore test whether copredication examples with Relatedness are judged *more* acceptable than equivalent examples without copredication. If there is no differential effect, then the effect of Relatedness is orthogonal to copredication.

4. EXPERIMENT SERIES 2: CONTROL BY THE IDENTITY METHOD

4.1. Pretests

To create the materials for our second experiment series we carried out further preparatory studies of which we again sketch two. The aim of the first was to establish which indicators unambiguously triggered either Event or Result readings. We first selected only those indicators to test which seemed to us introspectively to have a unique reading. We tested 12 different NP premodifying indicators, and 12 VP indicators, in equal proportions of Result and Event readings. The indicators were presented together with 12 nouns of three different types: 4 clear event nouns (e.g., *Gespräch* "conversation"), 4 clear object nouns (e.g., *Buch* "book"), and 4 examples from our list of nominalizations in *-ung* which can bear either Result or Event readings (e.g., *Auswertung* "analysis"). Participants thus saw these 12 nouns in four conditions in a 2×2 design with the factors Indicator Type (*Res*, *Ev*) and Indicator Place (NP, VP).

If an indicator has a unique reading, it should be judged good with the NP type which corresponds to this reading, and bad with the NP which does not

Table 4: Sample experimental items in pretest on forcing indicators.

Indic.	Noun types	Example stimulus
NP Res	Res/Ev/Amb	das/die wieder aufgetauchte Buch/Gespräch/Auswertung "the reappeared book/conversation/analysis"
NP Ev	Res/Ev/Amb	das/die kurzfristig vorverlegte Buch/Gespräch/Auswertung "the at.short.notice brought forward book/conversation/analysis"
VP Res	Res/Ev/Amb	Das/Die Buch/Gespräch/Auswertung ist wieder aufgetaucht. "The book/conversation/analysis has reappeared."
VP Ev	Res/Ev/Amb	Das/Die Buch/Gespräch/Auswertung musste vorverlegt werden. "The book/conversation/analysis had to be brought forward."

correspond. All indicators should be judged acceptable with all of our *-ung* nominalizations, since these are ambiguous in their reading. Indicators which do not produce the predicted results must be rejected. There are examples of the conditions in Table 4.

This experiment was carried out using the Thermometer Judgements method of gathering experimental relative judgements (Featherston, 2009), a development from Magnitude Estimation (Bard, Robertson, & Sorace, 1996). The method gathers introspective judgements relative to reference examples, which provide fixed points to anchor judgements and provide intersubjectivity. The task given to the participants is to express "how natural," in their own instant unreflected intuitions, example sentences are relative to the reference sentences. Experience has shown that informants are much more able to perceive and express relative acceptability than absolute acceptability (Anderson, 1992; Laming, 1997). Participants are instructed to give their judgements in numerical form, on a scale which has neither hard end points nor minimum division, but two fixed reference points, which bear the values 20 and 30, and which are anchored by example sentences. This method allows speakers the maximum possible freedom to express their intuitions without hindrance or deformation. It avoids the disadvantages of both zero points and multiples inherent in Magnitude Estimation, and the distortion of hard scale ends and fixed scale points associated with the traditional five or seven point scale (for details and further discussion see Featherston, 2008, 2009). Seventeen subjects participated in this study, all native speakers of German from the University of Tübingen.

The results showed that the participants generally shared the intuitions of the experimenters: the Result indicators are judged good with Result-type nouns and bad with Event-type nouns, while the reverse is true of the Event indicators. The results of the *-ung* nominalizations are much nearer the scores

Table 5: Sample experimental item in pretest on indicator interactions.

Sentence beginning (Res)	Sentence endings (all *Res*)
Das wiederaufgetauchte Paket ...	... ist beschädigt.
"The reappeared package ...	... is damaged."
	... besteht aus vielen Einzelteilen.
	... is made up of many separate parts."
	... muss ersetzt werden.
	... will have to be replaced."

of the conditions with matching indicator and noun type, though not quite as good as these. These findings were used to improve the quality of the experimental materials.

The aim of the final preparatory study was to ensure that the indicators in combination would produce the intended interpretation and not interact with each other in unintended ways so as to falsify the results of our study. Using the set of indicators confirmed as reliable in the previous preparatory experiment, we constructed 30 combinations of NP and VP indicators which both had the Event reading and 30 combinations which both had the Result reading. The indicator combinations were used with a small range of nouns with appropriate interpretations (Table 5).

Twenty-five participants recruited by e-mail from the University of Tübingen took part. All were native speakers of German. The experiment consisted of gathering their introspective judgements of the "naturalness" of the experimental materials. The methodology was the same as in the previous experiment. No clear cases were found where the indicators which had been found reliable on their own in previous studies interacted negatively with other indicators.

4.2. Main Study on Meaning Shift 2

While we have now developed linguistic materials which allow us to control for many irrelevant effects, we must now build experiments which credibly permit generalization to be drawn from their results. This requires us to use a sufficiently large number of exponents of the different conditions, so that the results cannot be argued to be specific to just those lexical items tested. We therefore proceed in two stages, because of the many values of the many parameters which must be varied.

Effectively it is not possible to test a wide range of nominalizations and a wide range of indicators simultaneously, since this makes the experiment impossibly large. We therefore test a larger number of indicators and a smaller number of nominalizations in this experiment to control for variation between indicators. In the following experiment we reverse this and test

fewer indicators with a larger set of nominalizations. We also test our hypotheses over the two versions of the experiment. In main study 2 we address the hypotheses:

i. both the sets of indicators and the set of nominalizations will produce homogeneous results,
ii. examples with matching indicators will be judged better than those with nonmatching indicators (contra underspecification),
iii. the *Ev*→*Res* order of indicators will be judged better than the reverse.

In this experiment we tested 10 *-ung* nominalizations together with a set of 10 different lexical indicators in the NP and 10 in the VP. The aim was to take a first step toward a quantification of the preference for matching indicators and dispreference for nonmatching indicators in materials which reasonably permit the results to be considered generalizable to the language as a whole. The nominalizations were presented in four conditions, two with matching indicators (*Res Res*, *Ev Ev*) and two with nonmatching indicators (*Res Ev*, *Ev Res*). The materials are listed in Table 6.

Table 6: Experimental materials in main study 2.

Event indicators in NP	**Nominalizations**	**Event indicators in VP**
1 zwei Stunden dauernde "lasting two hours"	Auswertung "analysis"	1 hat begonnen "has begun"
2 regelmäßig stattfindende "taking place regulary"	Bearbeitung "processing"	2 fand gestern statt "took place yesterday"
3 kurzfristig vorverlegte "brought forward at short notice"	Bemalung "painting"	3 wurde unterbrochen "was interrupted"
4 unterbrochene "interrupted"	Erzählung "narration"	4 dauerte lange "lasted a long time"
5 stundenlange "hours-long"	Gliederung "classification"	5 wurde fortgesetzt "was continued"
	Plakatierung	
Result indicators in NP	"postering"	**Result indicators in VP**
1 wieder aufgetauchte "reappeared"	Rahmung "framing"	1 ist beschädigt "is damaged"
2 verschwundene "disappeared"	Schnürung "stringing"	2 muss ersetzt werden "must be replaced"
3 beschädigte "damaged"	Übersetzung "translation"	3 liegt auf dem Tisch "is lying on the table"
4 verschenkte "given away"	Überweisung "money transfer"	4 ist wieder aufgetaucht "has reappeared"
5 aus mehreren Einzelteilen bestehende "consisting of many parts"		5 besteht aus mehreren Einzelteilen "consists of many parts"

Not all combinations of preindicators, nominalizations, and postindicators were used in the experiment. The full combinatory set would consist of 1,000 example sentences which is impractical to test and would contain many items with contradictory or tautological contents. The most acceptable 400 items were therefore selected, such that all nominalizations occurred equally often, they occurred equally often in each condition, and the conditions occurred equally often. On the methodological side we should note that this selection of the best combinations for testing is an application of the local optimum approach to control. The semantic and pragmatic contents of the lexis prevent the indicators and nominalizations being randomly assigned to each other in the materials. To this extent there is still a confound between the effects of the experimental conditions and the lexical exponents of them. A pure identity approach to control in such a study is simply not possible.

We divided the material into 10 lists, such that each experimental participant saw each nominalization four times and each condition 10 times but with a different combination of indicators. Participants also judged 10 standard items as fillers. The procedure in this experiment was Thermometer Judgements carried out on-line as before. The 40 participants were recruited from the participant volunteer list at Tübingen University by e-mail, and were paid for taking part.

The results of this study are illustrated in Figure 3. We carried out a repeated measures anova with analyses by subjects and by items (just for the record) on the results. There was a weak effect for the first indicator, and a strong effect for the matching of the indicators, but no significant differential effect for the nature of the second indicator, nor any significant interaction of these.

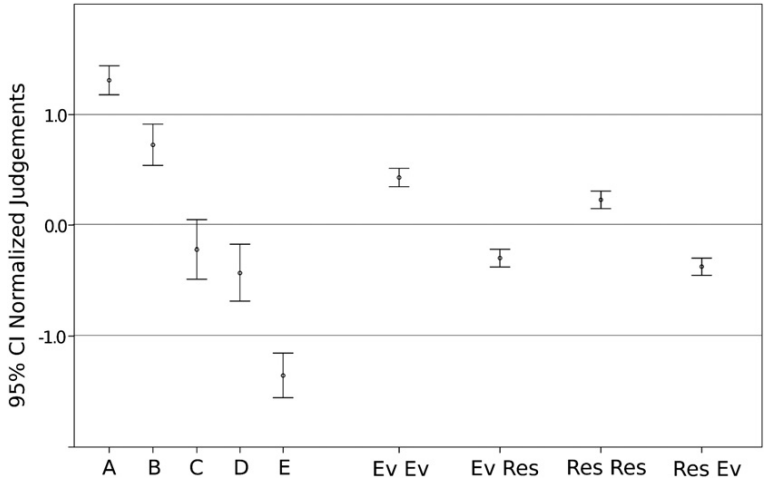

Figure 3: Results of main study 2 with standard comparison items.

On the left we see the results for the standard items which were developed for use in experimental syntax and which represent the full range of perceived acceptability, divided into five approximately equal parts. These standard items have been used in many experiments gathering perceived well-formedness and provide a comparison scale which allows an approximation to absolute well-formedness values (Featherston, 2009). Comparison with the standard items shows that the experimental sentences occupy the mid-range of the acceptability scale.

Looking at the experimental conditions, the results show that the prediction that matching indicators will be scored better is fulfilled; *Ev Ev* and *Res Res* are clearly better than *Ev Res* and *Res Ev*, as the statistics confirm (F_1 (1,39) = 124.8, $p_1 = 0.005$; F_2 (1,9) = 116.7, $p_2 < 0.001$). However, the expectation that the derivational ordering of nonmatching indicators (i.e., *Ev*→*Res*) would improve their rating is not confirmed, since the *Ev Res* condition and the *Res Ev* condition are very similarly scored. There is in fact a slight preference for the conditions with Event indicators in the early NP position, confirmed by the anova statistics for the factor First Indicator (F_1 (1,39) = 8.83, $p_1 = 0.005$; F_2 (1,9) = 5.05, $p_2 = 0.051$), but this is just a lack of complete homogeneity in the materials; the Event indicators in the NP must be slightly more natural. An effect of the derivational *Ev*→*Res* order of nonmatching indicators should reveal itself as an interaction of the factors First Indicator (*Ev*, *Res*) and Indicator Match (Matching, Nonmatching). There is however no such effect (all Fs < 2.5).

These results would suggest that our materials are sufficiently free of irrelevant effects to capture the difficulty of interpreting ambiguous nouns with contradictory indications from the linguistic context. These results should also be generalizable, at least over the indicators, since we have taken care to select them so as to exclude individual lexical effects. The consistent use of 20 indicators should suffice as a basis for a generalization about the effects of the indicators on the interpretation of the nouns. Our 10 *-ung* nominalizations by contrast are rather few to allow a generalization about the behaviour of this lexical group. We remedy this in main study 3.

Figures 4 and 5 show the experimental conditions distinguished by the lexical variants. The individual nominalizations in Figure 4 cluster quite satisfactorily, the only real inhomogeneity being in the *Ev Ev* condition, where some nominalizations pattern with the equivalent *Res Res* condition, while others score a little better. It is these better nominalizations alone which raise the *Ev Ev* result overall a little over the *Res Res* result. The two graphs in Figure 5 show the results by the NP indicators (on the left) and the VP indicators (on the right). The differences are fairly small, and we may be quite confident that the overall result is not the effect of just some of the lexical variants.

This data set may thus be seen as a vindication for the use of the "identity method" of materials creation, even in meaning-based studies. Recall that in our first study we used the "local optimum method," the lexical material in the indicator contexts being adapted to the meaning of the head noun. The results

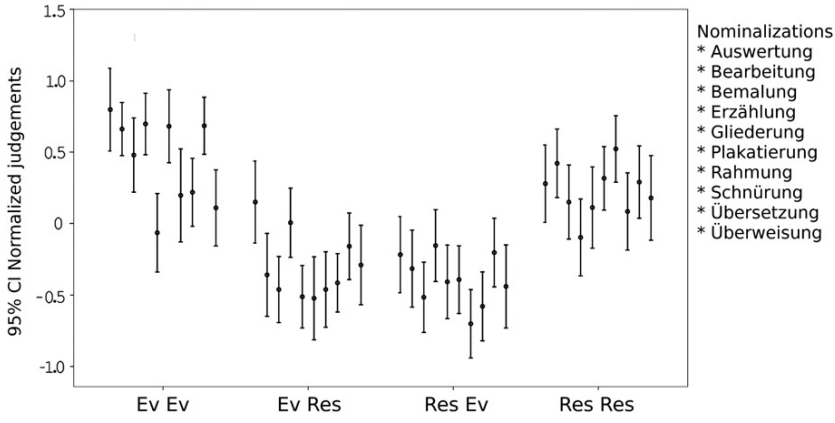

Figure 4: Main study 2: Results by conditions and nominalizations.

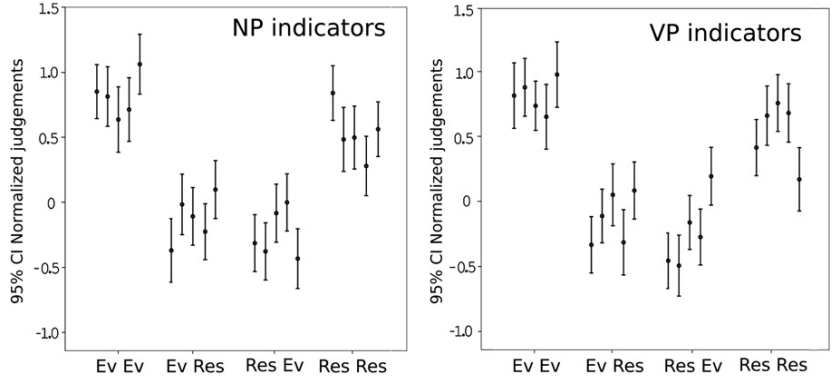

Figure 5: Main study 2: Results by conditions and indicators. The individual indicators appear as listed in Table 6.

of that study showed some quite clear differences among conditions which we would attribute to specificities of the lexis. These new results show a much more regular and systematic pattern of effects, which suggests that they are more generalizable and less dependent on the lexical items tested. We may therefore draw conclusions from them.

First, our use of standard comparison items with values from A to E in this last experiment allows us to state, with a degree of confidence, in absolute terms how unacceptable the conditions are (Featherston, 2009). The matching conditions approach the B value, which is fully acceptable, while the nonmatching examples are closest to the D standard item (unfortunately the C and D values are rather close in these results). Such examples are clearly awkward and flawed, and would

not be deliberately produced in this form, so it is questionable whether their forms should be regarded as "part of the language." They are however quite interpretable and permit an analysis within the structural constraints of the language; they are far from being nonsense strings.[2]

Second, this data thus replicates our previous tentative conclusion that the meaning shift does not appear as a specification of a previously underspecified representation. The drop in perceived acceptability (about one and a half steps, from B− to D+ on the standard items) looks too large for that (cf. the contrast between (4) and (5) above).

Third, these findings seem to demonstrate that there is no material effect of derivational direction in our process of interpretation. It is implicit in Nunberg (1995, 2004) that sense transfers have a direction, a point which is contested in Cruse (2004). Brandtner and von Heusinger (2010) had also wondered whether speaker-hearers might find a premodifying Event indicator with a postmodifying Result indicator easier to deal with, since this corresponds to both the derivational and causal ordering in our deverbal nominalizations. Intuitively, this seems quite plausible, but it does not appear in this data. One account of this intuition would be that an extension of lexical meaning from an Event reading to an additional Result reading is easier than an extension in the reverse direction. Put differently, it could be that the (Event → Result) effect holds, but as a cognitive preference leading to a generalization of the lexis of a language, rather than within individual items (cf. Traugott, 1990). If this were the case, our results here would not show the effect, even though it is true, since it would take the form of a greater difficulty in finding Event indicators that do not allow an extension to a Result interpretation than the reverse.

This experiment has verified and to an extent quantified the phenomenon of interpretation difficulty in case of mismatching indicators, but it does not provide evidence about the role of Relatedness in the phenomenon. Our next experiment addresses this.

5. MAIN STUDY 3

This final empirical study built upon the previous one by testing the same conditions with largely the same materials, but introduced the additional factor of Relatedness, which Nunberg suggests is a crucial variable (1995, 2004). We also rebalanced the lexical variants, using more nominalizations and fewer

[2]Some examples which exemplify the D grade of acceptability are in i.–iii. We offer no translation, as a translation of a degree of unacceptability is not possible. See Featherston (2009) for discussion.

i. Die Bergführer haben ihn einander als kompetenten Begleiter empfohlen.
ii. Wir lesen am liebsten die Süddeutsche, obwohl wir leben jetzt in Düsseldorf.
iii. Der Komponist hat dem neuen italienischen Tenor es zugemutet.

Table 7: Nominalizations in main study 3.

1	Absperrung	"barricading"	9	Isolierung	"insulation"
2	Auswertung	"analysis"	10	Kennzeichnung	"labeling"
3	Bearbeitung	"processing"	11	Neuerung	"renovation"
4	Bemalung	"painting"	12	Plakatierung	"postering"
5	Darstellung	"representation"	13	Rahmung	"framing"
6	Erzählung	"narration"	14	Übersetzung	"translation"
7	Garnierung	"garnishing"	15	Überweisung	"money transfer"
8	Gliederung	"classification"	16	Verpflegung	"catering"

indicators. Sixteen *-ung* nominalizations were tested which had been revealed to have the best balance of accessible Result and Event readings in our preparatory experiments (Table 7).

We used just two NP indicators each of Result and Event (for Event: *unterbrochene* "interrupted," *stundenlange* "hours-long"; for Result: *beschädigte* "damaged," *verschwundene* "disappeared") (Table 9). Since the previous studies demonstrated that our indicators were homogeneous in their effects, we are able to carry out this necessary reduction in the number of different NP indicators without risking the generalizability of the results. There was a little more variation in the VP indicators, for several reasons. First, these needed to demand Result or Event readings, but they also needed to either establish, or clearly not establish Relatedness with the first indicator and head noun. The specific pragmatics of the head noun and the need for Relatedness sometimes demanded slight changes in these VP indicators, sometimes lexical, sometimes grammatical. For example, sometimes the tense of the verb was varied or a time adverbial added. All VP indicators were variants of the seven Result and seven Event indicators listed here in Table 8.

On the basis of these materials we constructed 128 sentences made up of each of the 16 nominalizations in eight conditions in a $2 \times 2 \times 2$ design with the parameters NP Indicator (Result, Event), VP Indicator (Result, Event) and Relatedness (Related+, Related−). The sentences were divided into eight lists, such that each list contained each nominalization once and each condition twice. Fifteen standard items were added to each list as fillers (the same as in study 2 plus five more).

We should note one or two things here about the nature of the Relatedness relationship between the sentence parts. First, this Relatedness was most often instantiated by the suggestion of some sort of causal relation between the parts. If the head noun is qualified as being *beschädigt* ("damaged"), then it follows that it will be *repariert* or *erneuert* ("repaired" or "renovated"). Second, it was frequently sufficient to assert the connection with discourse signal words such as *nämlich* (something like: "you see") and *daraufhin* ("consequently"). Third, it proved to be less difficult than expected to produce examples which had

Table 8: VP indicators in main study 3.

VP Result indicators	VP Event indicators
… muss repariert/erneuert werden	… musste unterbrochen werden
"must be repaired/renewed"	"had to be interrupted"
… liegt auf dem Lastwagen/Tisch	… fand morgens statt
"is lying on the lorry/table"	"took place in the morning"
… war nämlich beschädigt	… wurde später fortgesetzt
"was you see damaged"	"was continued later on"
… ist wieder aufgetaucht	… wurde nicht beendet
"has reappeared"	"was not finished"
… besteht aus drei Teilen	… dauerte lange
"consists of three parts"	"took a long time"
… befindet sich im Haus	… hat begonnen
"is located in the house"	"has begun"
… wird nun endlich verpackt	… muss wiederholt werden
"is now at last being packed"	"must be repeated"

Table 9: Conditions in main study 3.

Ind_1	Ind_2	Rel.	Example
Res	Res	Rel+	Die beschädigte Absperrung muss repariert werden.
Res	Res	Rel−	Die beschädigte Absperrung liegt auf dem Lastwagen.
Res	Ev	Rel+	Die beschädigte Absperrung musste unterbrochen werden.
Res	Ev	Rel−	Die beschädigte Absperrung fand morgens statt.
Ev	Ev	Rel+	Die unterbrochene Absperrung wurde später fortgesetzt.
Ev	Ev	Rel−	Die unterbrochene Absperrung fand morgens statt.
Ev	Res	Rel+	Die unterbrochene Absperrung war nämlich beschädigt.
Ev	Res	Rel−	Die unterbrochene Absperrung liegt auf dem Lastwagen.

nonmatching indicators but a Relatedness relationship. Since the nonmatching indicators have been specifically selected to prevent a mutually compatible reading, one might expect that no Relatedness relationship between the parts should be possible. In fact, however, it seems to be quite feasible. Such examples as (25) and (26) seem to us to force incompatible readings of the parts, but at the same time indicate a causal or associative connection between the two. The reader is aware of the meaning shift, but is also conscious of a conceptual relationship.

(25) ?Die beschädigte Absperrung musste unterbrochen werden.
"The damaged barrier/barricading had to be interrupted."

(26) ?Die unterbrochene Absperrung war nämlich beschädigt.
"The interrupted barrier/barricading was, you see, damaged."

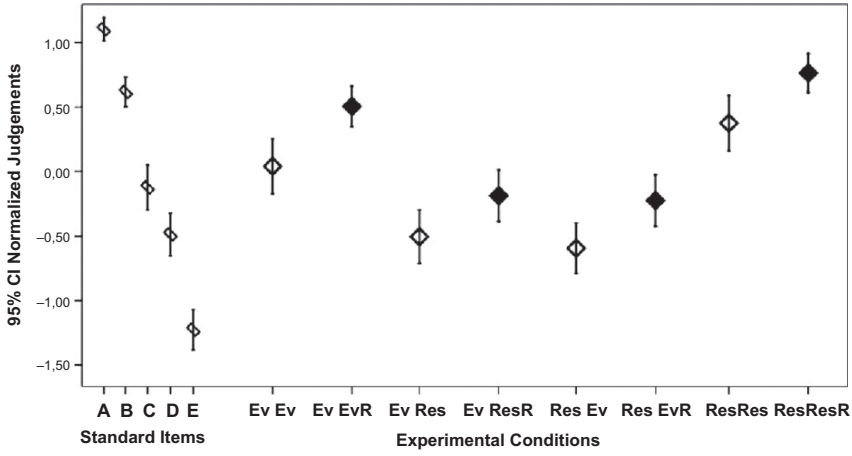

Figure 6: Main study 3: Results by condition plus standard comparison items.

The procedure in this experiment was, as before, Thermometer Judgements carried out on-line within Tübingen University. Thirty-six native speaker informants were recruited from the experiment participant volunteer list by e-mail, and paid for taking part. Informants were randomly assigned to the eight lists.

The results of the experiment as a whole are presented in Figure 6. These show our experimental conditions relative to the standard comparison items, which represent a scale of perceived well-formedness from A down to E. These results replicate the finding in main study 2 that all our experimental examples occupy the middle range between B and D. We may therefore confidently state that even the worst examples are not regarded by our participants as nonsense word strings, but only as marked expressions of the language, even in the worst cases. This is confirmed by the clear pattern in the experimental results: the eight conditions illustrated are systematically related to each other. Nonsense strings do not exhibit this.[3]

[3]Let us very briefly address a frequent question here: What do the *y*-axis values (−2 to +2) and the standard item values (A to E) mean? The numbers are *z*-scores, which means that the zero value is simply the mean of all values, +1 and −1 are one standard deviation up and down from that. Zero is in no sense a threshold of acceptability. The A to E scores on the other hand are an approximation to absolute grammaticality judgements (Featherston, 2009). The use of five points stems from the finding that informants can reliably give much finer judgements than just binary choices. This data therefore tells us *how much* better or worse one condition is than another. There is no meaningful threshold between good and bad. If it helps, readers could think of a C value as being a'?' judgement, a D value as a '??' judgement, and an E as a '*' judgement.

The conditions are given in the label under each error bar; *Ev Res* denotes a condition with an NP Event indicator and a VP Result, without Relatedness. The addition of R to the coding indicates Relatedness. We tested the results using the repeated measures anova procedure, applying the Huynh-Feldt correction when appropriate.[4]

There is no significant effect for the type of the first indicator, which suggests that our materials were well-balanced (both Fs < 2.5). There is, on the other hand, an effect of the second indicator (F_1 (1,40) $= 10.33$, $p_1 = 0.003$; F_2 (1,15) $= 4.84$, $p_2 < 0.044$), mainly due to the *Res Res* and *Res ResR* conditions being better than all the others. This is probably just a remnant inhomogeneity in the materials, for the previous experiment showed the opposite: a slight preference for *Ev Ev* over *Res Res*. There is naturally a strongly significant interaction of the effects of the two indicators, which represents the preference for indicators of matching types (F_1 (1,39) $= 130.2$, $p_1 < 0.001$; F_2 (1,15) $= 115.3$, $p_2 < 0.001$). There are no other significant interactions (all Fs < 2).

The most marked finding in this data set is that the values for the conditions with and without Relatedness differ sharply, but are systematically related. The anova reflects this with a significant effect for the factor Relatedness (F_1 (1,39) $= 29.67$, $p_1 < 0.001$; F_2 (1,15) $= 57.45$, $p_2 < 0.001$). There is however no sign of an interaction of Relatedness and any other factor, so the effect of Relatedness is apparently constant.

6. DISCUSSION: THE PHENOMENON OF COPREDICATION

Let us first note that this data further replicates the finding that items with matching indicators are perceived to be clearly more acceptable than those with nonmatching indicators. Furthermore, since we have now tested this both with larger numbers of indicators in the previous experiment and larger numbers of nominalizations in this experiment, we are approaching a supported generalization.

Given that our findings confirm the data pattern that the literature assumes, we may be sure that our results reflect the linguistic phenomenon Nunberg addresses. It is therefore interesting to observe what our results tell us about his account. Nunberg suggests that a Relatedness relation (in fact, a salient "functional relation" and a "noteworthy" relation; 1995, p. 114) is required for meaning shift to be acceptable, and supports this with good examples where such relations are present and contrasting bad examples where they are absent. Our studies have found the facilitating effects of matching indicators and Relatedness that his account predicts. To that extent his claims are supported.

[4]We again report by item analyses purely for the record.

But our study also tested whether these effects are a condition of the acceptability of meaning shift, or whether the improvement in acceptability is an independent effect. The results show clearly that there is no link between Relatedness and the acceptability of the meaning shift; the improvement in perceived acceptability is just as large in the conditions with matching indicators and therefore no necessary meaning shift, such as *Res Res* and *Ev Ev*, as it is in the conditions with nonmatching indicators, *Res Ev* and *Ev Res*. So Nunberg's acceptable examples of copredication are indeed only possible when they have some internal coherence, but this is a condition on acceptability or felicity more generally, not a condition on copredication.

This conclusion is supported by a little more detailed consideration of what made up Relatedness in our example sentences and how it triggered the facilitating effect. We have offered no definition (nor does Nunberg), but we found in pilot tests that other people shared our own intuitions about which examples exhibited it. Looking at the examples more closely, we may distinguish two main relationships between the parts which seem to yield the subjective impression of Relatedness. The first is a fairly simple association: there is often a lexical or conceptual contiguity between the two parts such as in (27) between *gestrig* ("yesterday") and *heute* ("today").

(27) Die gestrige Überweisung wird erst heute verbucht.
"Yesterday's money transfer will be entered in the books only today."

The second relationship is that of causality: one part is in some way the reason for the other. In (28) it seems plausible that a long exposition should have to be interrupted, the modal *musste* implying that the interruption was dependent upon the duration.

(28) Die stundenlange Darstellung musste unterbrochen werden.
"The hours-long exposition had to be interrupted."

This cannot be lexical priming because the causality is often not shown by the content words in the indicators, rather it is implied by markers of temporal or causal structure (e.g., *daraufhin* "consequently," *wieder* "back again" in the sense of restitution). In example (29) it is the discourse marker *nämlich* that tells us that an explanation is coming (and which we have glossed as "you see") and makes the processing of the following text easier, because we know what its function in the discourse will be.

(29) Die unterbrochene Absperrung war nämlich beschädigt.
"The interrupted barricade/barricading was damaged, you see."

The felicitation of examples with Relatedness might therefore be attributed to "discourse priming." More generally, we can think of the effect of Relatedness as being one of coherence (e.g., Kehler, 2002). Since it is such a general effect,

it is not surprising that we find it in all types of example sentences, not only those with copredication.

Nunberg's (1995, 2004) second major claim stands unaffected by these results: he argues that, in examples such as (30), it is the predicate which undergoes meaning shift, rather than the subject, so that (30a) is the interpretation, not (30b).

(30) I am parked on a pedestrian crossing.
  a. "I am [the driver of a car which is] parked on a pedestrian crossing."
  b. "[My car] is parked on a pedestrian crossing."

We have found no way of experimentally testing this question, but we make a number of comments in Weiland et al. (2010). One approach to the change in reference in this example makes use of the concept of a Driver as a sort: an action unit consisting of a person and a vehicle. Another approach might be to assume that (30) is interpreted not as one proposition, but as two propositions, as in (31). The references to the speaker and the vehicle can remain disjoint.

(31) Proposition 1: I am the driver of a vehicle.
  Proposition 2: This vehicle is parked on a pedestrian crossing.

It is interesting to note that these two propositions closely resemble the separate clauses of Nunberg's own suggestion of the form of the enriched predicate (30a). The two-clause solution is thus not so far from Nunberg's own.

7. METHODOLOGICAL CONCLUSIONS

This series of studies revealed quite how difficult it can be to obtain firm judgement data on a topic such as the one addressed in this chapter. Most previous work using experimentally obtained judgements has been in experimental syntax, and the research aims have been reducible to generalizations about structure. In these the effects of lexis are systematically controlled for and thus excluded. The design of this study however included specifications about the lexical content of the materials, and the experimental conditions also have meaning-related components. Such a study places stringent constraints upon the materials.

First, it requires us to establish in multiple pretests which lexical materials fulfill the specifications; for meaning-related requirements are much less easily established than form-related requirements. That a particular lexical set contains only nouns which are matched for length, frequency, and phonotactics can be verified in only a few minutes; whether a nominalization has both Event and Result readings which are equally accessible can only be established in a

pretest. This is of course more work, but it has other implications too: the narrow preselection threatens the claim of the results eventually gained to be generalizable, which reduces their validity considerably.

The strength of the experimental method is that the use of control and random selection permit the findings to be assumed to apply much more widely than just the set tested. The nominalizations which we tested were only a selection of all the nominalizations with *-ung* in the German language, but the research question is a larger one. The nominalizations we test are supposed to be representative of the complete lexical set of this form, but they are also just one example of a wider derivational process, to which our results should ideally be applicable. So can we generalize from our results?

It is, for example, tempting to suggest on the basis of the results in Figure 6, which show the *Res Res* and *Res ResR* conditions to be judged better than all the others, that nominalizations in *-ung* prefer a Result interpretation. But no such conclusion is possible. We carried out extensive pretesting to find those lexical items which allowed the Event and the Result readings most equally readily. So all that can be asserted with confidence is that this set of 16 nominalizations is preferred with this reading. Lexical sets do not generalize unless they are randomly selected for the parameters of interest; for our materials this can hardly be asserted. Control of materials in an experiment like this is thus a double-edged sword: too little control would permit the result to be falsified by outliers, but too much control risks the selection of certain characteristics and features which are not representative of the full set.

We may therefore draw two perhaps surprising conclusions. Rather than being an unalloyed good thing, control of the linguistic materials must be regarded as being in a trade-off relationship with random selection: the more control, the less random selection, and vice versa. Both are necessary features of the linguistic materials to permit generalization. In many cases this will not matter too much because there will be both sufficient control and sufficient random selection, but this experimental series demonstrates that it is possible for hypotheses to be so specific and constrained from so many different directions that they verge on being unverifiable, although the author has illustrated the claim with examples. For the constraints on the materials may limit the possible lexical exemplars to just about exactly those which illustrate the claim. If this is so, then the claim lacks generalizability and therefore wider relevance.

This highlights one reason why it is useful for linguists to test their claims experimentally: the process of gathering multiple examples can bring such problems to light. If no more or very few more examples of a phenomenon can be found, the effect apparently lacks the wider scope the linguist had imagined for it. Not only from experimental results, therefore, but also from the experiment design and construction process are insights into the applicability of hypotheses gained.

Another issue concerning control which we should like to touch upon here is the method of control applied. In the first version of these experiments we employed the "local optimum" method of experimental materials construction. In this approach, each lexical variant of the materials is made as acceptable as possible, rather than as similar as possible to the others. This approach comes at the price of permitting a greater degree of lexical and content variation over lexical variants. It is however sometimes the best available method, when other constraints on the lexical form make it difficult to produce lexical variants which can remain constant across variants. In this study the nominalizations were preselected first by morphological form (*-ung*) but further by meaning, since they must equally well bear Event and Result readings. These two factors already very narrowly restrict the set of nouns that we can use in our lexicalizations, and other examples drop out because of their low frequency. Inevitably, the remaining few refer to very different things; one cannot say the same things about an analysis ("Auswertung"), catering ("Verpflegung"), and insulation ("Isolierung"). In such a case, it can be best to simply make each lexical variant as natural as possible, on the assumption that they will thus be matched, since they are all fully natural.

We adopted this local optimum method of materials creation in our first experiment. However, detailed analysis of the results of our first experiment showed that there were significant differences between the lexical variants, which severely affected the overall result. We therefore chose to switch to the "identity" method of constructing materials, again carrying out preparatory experiments. This approach finally proved itself to be more adequate.

The desire to be able to claim generalizability for both the indicator effects and the behaviour of the nominalizations meant that both had to be tested in considerable numbers. Constraints on the feasible size of experiments meant that it was only possible to test the full variety of indicators with a reduced range of nominalizations and vice versa. Some adjustments of the materials were necessary, both in order to make feasible combinations of indicators and nominalizations, but also in order to superimpose on these materials the contrast of Relatedness and the lack of it.

The best achievable materials therefore started from the fully controlled identity method, but nevertheless had considerable aspects of the local optimum method. This may be the best that can be obtained in such a study. In spite of these compromises in the materials, the quality of the results of these studies is high, as can be seen in the reasonable clustering of the scores of the individual item components in Figures 4 and 5, and the clearly systematic response in the eight conditions in Figure 6. We therefore finish on a positive note: although questions of interpretation are notoriously hard to pin down, our studies have been successful in capturing the linguistic phenomena under discussion and have yielded some insights into their analysis. And as often, the experiment building provided as much clarification as the result.

ACKNOWLEDGMENTS

This work took place in the context of the SFB 732 project B1 "The formation and interpretation of derived nominals" supported by the Deutsche Forschungsgemeinschaft. Thanks are due to Regine Brandtner and Edgar Onea for discussion and advice. All remaining weaknesses are our own.

REFERENCES

Anderson, N. (1992). Integration psychophysics and cognition. In: D. Algom (Ed.), *Psychophysical approaches to cognition* (pp. 13–113). Amsterdam: North Holland.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge: CUP.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*(1), 32–68.

Brandtner, R. (2011). *Deverbal nominals in context: Meaning variation and copredication*. Ph.D. dissertation, University of Stuttgart.

Brandtner, R., & von Heusinger, K. (2010). Nominalization in context-conflicting readings and predicate transfer. In: A. Alexiadou & M. Rathert (Eds.), *Nominalizations across languages and frameworks* (pp. 25–50), Berlin: de Gruyter.

Bresnan, J., & Nikitina, T. (2009). The gradience of the dative alternation. In: L. Uyechi & L. Hee Wee (Eds.), *Reality exploration and discovery: Pattern interaction in language and life* (pp. 161–184), Stanford: CSLI.

Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, *12*, 15–67.

Cruse, A. (2004). Lexical facets and metonymy. *Ilha do Desterro*, *47*, 73–96.

Ehrich, V., & Rapp, I. (2000). Sortale Bedeutung und Argumentstruktur. *Zeitschrift für Sprachwissenschaft*, *19*, 245–303.

Featherston, S. (2008). Thermometer judgements as linguistic evidence. In: C. M. Riehl & A. Rothe (Eds.), *Evidenz* (pp. 69–89). Aachen: Shaker.

Featherston, S. (2009). A scale for measuring well-formedness; Why linguistics needs boiling and freezing points. In: S. Featherston & S. Winkler (Eds.), *The fruits of empirical linguistics. Volume 1: Process* (pp. 47–74), Berlin: de Gruyter.

Frazier, L., & Fodor, J. (1978). The sausage machine: A new two stage parsing model. *Cognition*, *6*, 291–325.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford: CSLI.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavioral Research Methods*, *41*(1), 1–12.

Laming, D. (1997). *The measurement of sensation*. Oxford: OUP.

Nunberg, G. (1995). Transfers of meaning. *Journal of semantics*, *12*, 109–132.

Nunberg, G. (2004). The pragmatics of deferred interpretation. In: L. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 344–364). Oxford: Blackwell.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Traugott, E. (1990). From less to more situated in language: The unidirectionality of semantic change. In: S. Adamson, V. Law, N. Vincent & S. Wright (Eds.), *Papers from the Fifth International Conference on English Historical Linguistics* (pp. 496–517). Amsterdam: Benjamin.

Weiland, H., Featherston, S., & von Heusinger, K. (2010). Conditions on meaning shift: An empirical investigation. Manuscript, Universities of Stuttgart, Tübingen.