

# Prominence and semantics in English compounds

Dominic Schmitz, Melanie J. Bell & Ingo Plag

May 13, 2026

## Abstract

English noun-noun compounds can be left-prominent or right-prominent. Previous accounts of this variation found that aspects of the compounds' semantics play some role in the distribution of the two prominence patterns. The nature of the relation between semantics and prosody in compounds has remained largely elusive, however. This paper explores the relation between compound semantics and compound prosody starting from the hypothesis that observed effects emerge from a language system originating in the speaker's experience, through a process of discriminative learning. Using the Boston University Radio Speech Corpus, we created discriminative learning networks that map form onto meaning, modelling comprehension, and meaning onto form, modelling production. Context-sensitive, token-based word embeddings were created using BERT. Pitch contours were derived by using generalised additive models. The statistical analysis of the pitch contours lends support to the existence of the two stress patterns advocated in large parts of the literature. We demonstrate that it is to some extent possible to predict the meaning of a compound based on its pitch contours, and, to a lesser extent, to predict the pitch contour, given the meaning. The theoretical implications of these findings are discussed.

**Keywords:** compounds, prominence, stress, discriminative learning

# 1 Introduction

The prominence relation between the two constituents in English noun-noun compounds is variable. According to pertinent reference works (e.g. Bauer et al. 2013, chapter 19), there are compounds that are left-stressed, e.g. *Oxford Street* and *opera singer*, while others have right stress, e.g. *Oxford Road* and *summer dress*<sup>1</sup>. Empirical studies have shown that a third or more of the compound tokens in naturally occurring speech are right-stressed (e.g. Bell & Plag 2012; Kunter 2011; Plag 2010).

The variability of stress assignment in noun-noun compounds has been a test case in the debate about the nature and role of symbolic rules, associative networks and analogical mechanisms in the organization of language. This interest has been fed by an increasing awareness even in generative linguistics of the fuzziness, semi-regularity and irregularity of many phenomena on all levels of linguistic description. Solving the puzzle of variable compound stress thus has important implications for the question of how linguistic knowledge is to be represented in a descriptively and explanatorily adequate model of grammar and lexicon.

Studies of the acoustic properties of compounds (e.g. Farnetani et al. 1988; Plag 2006; Kunter & Plag 2007; Kunter 2011) have demonstrated that pitch and intensity are the most important acoustic correlates of compound stress. Phonologically, the different stress patterns (left or right) have been argued to manifest themselves by the presence or absence of pitch accents on the constituents (e.g. Kunter 2011). Left-stressed compound tokens have only one pitch accent, which is realised on the most prominent syllable of their left constituent, whereas right-stressed tokens have a pitch accent on the prominent syllable of both constituents. The pitch accents in turn are realised primarily through pitch. Roughly speaking, left-stressed compounds have higher pitch in the left constituent, while right-stressed compounds show only small differences in pitch height between the two constituents. In addition, there are differences in the pitch contours. Figure 1 shows the stylised contours as given by (Kunter, 2011, 95), which show a rising and falling contour for left-stressed compounds, and two falling contours for right-stressed compounds.



Figure 1: Stylised pitch contours of left- and right-stressed compounds, adapted from Kunter (2011:95)

Variation in English compound stress has been investigated in detail in a number of large empirical studies

<sup>1</sup>We use the terms ‘prominence’ and ‘stress’ interchangeably, as nothing hinges on this distinction in the context of this paper.

(e.g. Arndt-Lappe 2011; Bell 2015a,b; Bell & Plag 2012, 2013; Kunter 2011; Kunter & Plag 2007; Plag 2006; Plag et al. 2007, 2008; Plag 2010). This body of work has demonstrated that various factors at different levels of representation correlate with stress patterns in English compounds:

- Compounds with the same left or right constituent tend to have the same stress pattern. This is often interpreted as an analogical effect.
- A constituent is less likely to be stressed if it occurs in the same position, i.e. as head (N1) or as modifier (N2), in a large number of compounds ('family size effect').
- Greater semantic specificity increases a constituent's chance of stress.
- Certain semantic classes of constituents (e.g. 'N1 is a location') and certain semantic relations (e.g. 'N2 is made of N1', 'N1 is for N2') have a tendency to be associated with particular stress patterns. For example, proper nouns in N1 position, e.g. *Lynx helicopter*, and the 'is made of' relation, e.g. *silk shirt*, are associated with right stress.
- More frequent and lexicalised compounds tend to have left stress.
- Longer compounds, in terms of syllable count, tend to be right-stressed.

Compound stress patterns vary with dialect, but even within regional varieties we find stress doublets. For instance, the OED (2022) lists both *íce cream* and *ice créam* for both British and U.S. English, but in different orders, presumably to indicate that the right-stressed variant is more prevalent in Britain, and the left-stressed one in the U.S. For *weekend*, the OED lists both stress variants for Britain, but only the left-stressed one for the U.S. In his discussion of the phonetic manifestations of compound stress, Kunter (2011) notes that different varieties of English seem to differ not only in the position of stress but also in the way they implement it phonetically.

Over and above dialectal and other sociolinguistic differences, there is considerable stress variation between and within individual speakers, which is not straightforwardly explained by contextual effects such as contrastive stress. For instance, in Bell & Plag's (2012) sample using data from speakers of British English, 37 percent of the 864 compound types exhibit inter-speaker variation in stress. Furthermore, some compounds show more stress variability than others. In her study of variable compounds, Bell (2015b) finds that variation is more likely when the different factors influencing stress placement exert conflicting pressures. Given that semantics plays a role in stress assignment, some of the between-speaker variation might also be due to differing individual conceptualizations, especially for low frequency compounds, since novel compounds are open to a very wide range of interpretations (Schäfer & Bell, 2020).

The empirical findings on compound stress also bring up other important questions. There is no simple mapping of acoustics to stress perception, and it turns out that there is also considerable variation in people's

perception of stress (Kunter, 2010, 2011). This raises some rather fundamental concerns about the reality of categorical binary stress patterns and about how such stress patterns have been dealt with in linguistic theory (see also Schmitz et al. 2026). On the semantic side, we find similar problems. Determining the semantics of compounds is far from trivial, and even trained raters vary in their coding of presumably important semantic properties, even for institutionalised compounds with established meanings (Ó Séaghdha, 2008).

To summarise the results of the existing empirical work, compound stress is not governed by deterministic rules but depends probabilistically on the distribution of properties across other compounds in the speaker’s mental lexicon. However, the large amount of variation and the multitude of factors that have a say in the distribution of stress in compounds raise difficult and very important questions: How do speakers actually make use of these factors? How are these factors and their respective weights represented in the speaker, and how does the speaker learn and know how to apply these factors? What kind of linguistic architecture is able to cope with the complexities of the variability?

The present paper explores the relation between compound semantics and compound prosody starting from the hypothesis that the observed effects emerge from a language system that originates in the speaker’s experience, through a process of discriminative learning, in which forms are mapped onto meanings (comprehension), and meanings onto forms (production). We will test this hypothesis using the computational architecture of discriminative learning theory, as implemented in the ‘Discriminative Lexicon’ model developed by Harald Baayen and colleagues (e.g. Baayen et al. 2018a, 2019; Chuang & Baayen 2021).

We model compound semantics directly from the acoustics of the speech signal, and the acoustics directly from the semantics. This involves two important methodological innovations. First, we move away from problematic perceived or rated prominence and use instead the most important part of the acoustic signal, i.e. pitch contours. Second, we use distributional semantic vectors to represent the meaning of compounds instead of focusing on the influence of individual semantic properties, such as certain semantic relations. It is demonstrated that, using linear discriminative learning (LDL) networks, it is indeed possible (to some extent) to predict a compound’s meaning from its pitch contours, and its pitch contour from the compound’s meaning. This has important theoretical implications, which we will discuss in section 6.

After giving an overview of the issues involved in the relation of prominence and semantics in English noun-noun compounds, we will first investigate the phonetic properties of the compounds in our data set and their phonological interpretation (study 1). We will then explore the distributional semantic space of the compounds with regard to important semantic properties discussed in the literature (study 2). Finally, we will map the pitch contours onto semantics in a linear discriminative learning network (study 3).

## 2 Compound prominence and semantics

### 2.1 Compound prominence

As mentioned in Section 1, compound prominence has largely been discussed in terms of phonological categories, i.e. leftward and rightward stress. Sometimes a third category is recognised, ‘level stress’. These categories are based on the intuitions of researchers and usually not questioned. However, as has been repeatedly observed (e.g. by Plag, 2006; Plag et al., 2008; Kunter, 2010), the assignment of compounds to the two stress categories left and right is often very difficult for raters. Kunter (2010), for instance, discovered that almost half of his participants are not able to provide consistent judgments. Variability in the prominence ratings is particularly strong among the whole group of participants if the compound is generally perceived as right-prominent. Kunter argues that left prominence is easier to perceive and classify because the raters need to compare accented with unaccented material, i.e. between two different phonological categories. With right-prominent compounds, the comparison involves two elements that belong to the same phonological category (i.e. two pitch accents). In general, discrimination accuracy in within-category comparisons is worse if the distinction is to be perceived categorically (Liberman et al., 1957; Repp, 1984).

A closer look at the acoustic correlates of compound stress reveals that stress ratings correlate with quite a few phonetic parameters. Kunter & Plag (2007); Farnetani et al. (1988); Kunter (2011); Plag et al. (2008); Ingram & Nguyen (2007) all find that pitch is a very good cue for prominence. In addition, intensity, duration, spectral tilt of the left constituent, and average absolute pitch slope in the right constituent are indicative of the prominence relationship between the two constituents. If the left constituent is more prominent than the right constituent, it will have a higher pitch, a higher intensity and a longer duration than the right constituent. Prominent left constituents tend to have a flat spectral balance, and a large pitch movement on the right constituent goes together with a more rightward prominence perception.

Right-prominent compounds feature level pitch and intensity across the two constituents. The latter fact raises the question of why a level pitch and intensity is interpreted phonologically as rightward stress. This is the result of the declination effect (Collier, 1975): F0 and intensity decrease steadily across an utterance, irrespective of the overall intonation pattern. In phonetic studies, it has been shown (e.g. in Gussenhoven et al. 1997; Gussenhoven & Rietveld 1988; Ladd et al. 1994; Rietveld & Gussenhoven 1985) that in a sequence of two pitch peaks, the second peak is perceived to be as prominent as the first peak even if the second pitch excursion is lower than the first one, which accounts for the claim that the nuclear accent in an intonational phrase is perceived as most prominent.

Compound stress involves a prominence relationship between two stressed syllables, and hence resem-

bles what is called primary stress and secondary stress in non-compound words, such as *hýphenàte* vs. *hýphenátion*. There is only one study that has systematically investigated the phonetics of primary and secondary stress in English, Plag et al. (2011). These authors found that the position of primary and secondary stress correlates with F0, intensity, duration, spectral balance in both constituents, and pitch slope in the left position. Based on their analysis, the authors come to the conclusion that (in accented positions) left-prominent words are phonologically characterised by one pitch accent, while right-prominent words have two pitch accents, one on each of the two stressed syllables. The phonetic correlates and phonological interpretation of primary and secondary stress in non-compounds are thus analogous to the correlates and phonological interpretation in compounds.

Figure 2 illustrates the pitch contours of the two compounds *campáign pròmise* (left-stressed) and *hòme phóne* (right-stressed) from the Boston University Radio Speech Corpus (Ostendorf et al., 1996). One can clearly see one fall (in the left panel) vs. two falls (in the right panel), which indicate the difference between one pitch accent (left stress) and two pitch accents (right stress).

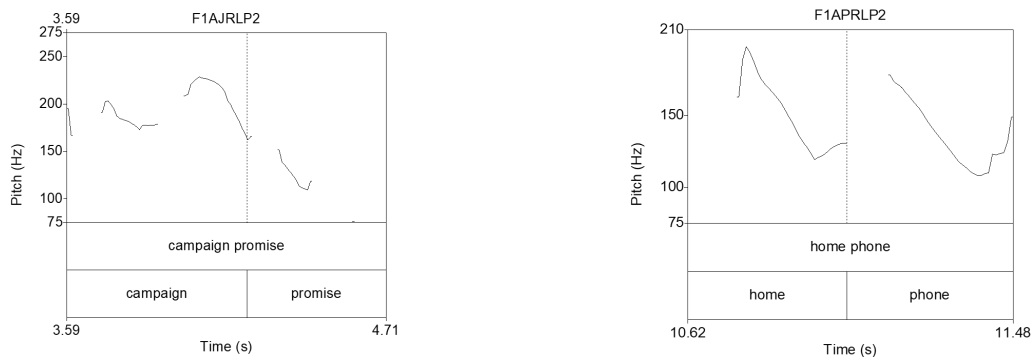


Figure 2: Pitch contours for *campaign promise* (left panel) and *homephone* (right panel)

Phonetic analyses of prominence have traditionally used single measurements in certain intervals, for example, pitch maxima, or mean pitch, instead of whole contours. One reason for this is the assumption that phonologically, tonal targets are identified that abstract away from the details of the contour and focus instead on peaks, troughs or plateaus that are deemed phonologically relevant. Reducing the rich signal encapsulated in the contour, may, however, run the risk of disregarding the systematic variation of the contour that is potentially highly informative for detecting prominence.

Listeners are attuned to the shape of the pitch contour, and perceive those elements as particularly prominent during which the pitch contour shows particular events. This link between pitch and prominence has been supported by the findings in studies such as Rietveld & Gussenhovent (1985) and Terken (1997): syllables received increasingly higher prominence ratings with increasing pitch excursion sizes. Rietveld

& Gussenhovent (1985) have shown that for Dutch, a pitch excursion of 1.5 semitones is a sufficient cue to prominence. Terken (1997) accordingly concludes that higher prominence ratings indeed appear to be proportional with the size of pitch excursions.

The direction of movement may also be important for recognizing prominence. For instance, Terken & Hermes (2000) demonstrated that a fall is more readily perceived as more prominent than a rise of the same excursion size. One-dimensional measures like peaks or averages cannot reflect movement in the contour. Furthermore, as shown by Arvaniti et al. (1998), tonal targets may not be aligned properly with the accented syllable. A local measure of the wrong interval may therefore miss the relevant pitch information altogether.

Kösling et al. (2013) make use of the whole pitch contour to investigate the prominence patterns of triconstituent compounds (such as *hay fever treatment* or *prisoner community service*) and show that the analysis of pitch contours can provide important evidence for an accent-based account of compound prominence. The pitch contours of Kösling et al. (2013)'s compounds showed a rather clear general downward trend (i.e. declination), which was sometimes interrupted by a plateau or small occasional rises. The authors interpreted a high start of a pitch contour as an indication of a pitch accent. Additional pitch accents (if any) were identified by rises of at least one semitone or by clear falls.

Schmitz et al. (2026) investigate the role of context on the prosody of compounds, also making use of the compounds' pitch contours. In their implementation of k-means clustering two clusters is the optimal number of clusters. One of the clusters clearly comprises left-stressed compounds, while other cluster of contours is not so readily interpretable as right-stressed.

In the present study, we will also make use of pitch contours instead of one-dimensional single measurements.

## 2.2 Compound semantics

Empirical studies of the relationship between compound prominence and semantics investigated effects of argument structure (which includes syntactic aspects), the semantic categories of the constituents or the whole compound, the semantic relation between constituents, and the semantic specificity of the second constituent (N2) (see, e.g. Bauer et al. 2013, ch. 20 for an overview).

Giegerich (2004) hypothesised that complement-head structures like *trúck driver* are left-stressed, while modifier-head structures such as *steel brídge* are generally right-stressed (unless lexicalised with left stress). Plag et al. (2007, 2008) tested this claim with different data sets, and found the expected left stress effect of argument structure only for compounds whose head features the suffix *-er* (e.g. *law makers*, *screw driver*).

Exploring other claims from the literature (e.g. Fudge 1984, 144ff; Liberman & Sproat 1992; Zwicky

1986), Plag et al. (2007, 2008) found that the following semantic properties of constituents or compounds exert some influence: ‘N1 refers to a period or point in time’, ‘N2 is a geographical term’, ‘N1 is a proper noun’, ‘N1 and N2 form a proper noun’ and ‘N1 and N2 form a left-headed compound’ tend towards rightward stress.

The nature of the semantic relations between the constituents have been investigated in many publications (often with a psycholinguistic orientation, see Benjamin & Schmidtke 2023 for a recent study), and there is evidence that these relations also have a bearing on stress. Using 18 relations gleaned from the literature, Plag et al. (2008) demonstrate that seven relations play a role in stress assignment: ‘N1 has N2’, ‘N2 is made of N1’, ‘N1 is N2’, ‘N2 is located at N1’, ‘N2 during N1’, ‘N2 is named after N1’ favor right stress, while ‘N2 uses N1’ tends towards left stress. Using CELEX (Baayen et al., 1996), Plag et al. (2007) also find effects of semantic categories and semantic relations that largely overlap with the ones found in Plag et al. (2008). Bell (2015b) proposes that these relations can be conceived of as particular instantiations of the broader category of ‘basic relations’, proposed much earlier (for German compounds) by Fanselow (2011).

Bell & Plag (2012, 2013) investigated the role of informativity for compound stress. They coded semantic specificity as one aspect (and measure) of informativity and their data revealed that semantically more specific right constituents are more likely to receive an accent, which in turn means rightward stress.

To summarise, there is a lot of evidence for a link between semantics and prominence in compounds. However, the nature of this relationship is rather unclear, for methodological and theoretical reasons. As has been pointed out repeatedly, most of the semantic categories and relations are rather ill-defined, which makes it hard to classify compounds unambiguously in these taxonomies. Even trained raters vary in their coding of compound semantics, even for institutionalised compounds with established meanings (Ó Séaghdha, 2008). Furthermore, compounds and their constituents are often polysemous (e.g. Schäfer & Bell 2020), such that multiple codings are frequently necessary. At a theoretical level, the definition, number and selection of semantic properties is quite problematic. Researchers have addressed this issue pragmatically (but theoretically quite unsatisfactorily), by using properties that others have used before, like Levi (1978)’s semantic relations.

For these reasons, the discrete semantic properties coded in empirical studies of compounds seem rather questionable idealisations of the complexities of compound semantics. Furthermore, it is unclear why certain semantic categories do not seem to correlate with a particular stress pattern at all, while others do (see, for example, Plag et al. 2008). And for those categories or relations that do show a correlation, it is unclear why they should prefer to correlate with one of the stress patterns, e.g. rightward stress, and not with the other stress pattern. Why would, for example, compounds exhibiting the relation ‘N2 is made of N1’ prefer rightward stress? Given all these problems, it can be stated that we still do not have a good understanding

of compound semantics, and, as a consequence, its bearing on compound stress.

More recently, scholars have begun to use an alternative approach to compound meaning, distributional semantics. Numerous studies have shown that using embeddings of compounds and their constituents can enhance our understanding of compound semantics. For instance, there are studies that successfully employed embeddings for the whole compound and its constituents to predict the semantic transparency of compounds (e.g. Alipoormolabashi & Schulte im Walde 2020; Schulte im Walde et al. 2016). Other studies have tried to incorporate compositionality into their models. Algebraic models use mathematical functions like addition, pointwise multiplication, weighted addition, dilation, or tensor products to create embeddings for compounds on the basis of their constituent embeddings (see Mitchell & Lapata 2010 for some discussion). The CAOSS model (Marelli et al., 2017)) derives representations of compound meanings that take into account the role of the constituent as either modifier or head. The validity of such compositional models has been tested by using the embeddings (or measures derived from them, such as cosine similarities) to predict reaction times in behavioral experiments, with impressive success (e.g. Günther et al. 2020; Günther & Marelli 2021; Petilli et al. 2019). With their CAOSS-derived vectors, Günther & Marelli (2022) were able to quite accurately predict the relational interpretations for familiar as well as novel compounds as provided by human participants, and additional qualitative explorations showed the usefulness of the results also from a theoretical-linguistic point of view. The treatment of particular compounds by the system was in conformity with traditional semantic considerations based on the categories and relations discussed above. This means that embeddings contain the semantic information that has traditionally been described in terms of discrete properties.

Notably, available studies of compound embeddings mostly use established compounds that are spelled as one word. Productively formed compounds are, however, often spelled as two words, and they are of very low frequency, which raises the problem of how to obtain reliable vectors for them. Another problem is the polysemy of the two constituents. Compound constituents (like other nouns) are often polysemous, and particular compounds may involve quite different meanings of the same constituent. Consider, for instance, the compounds *chainsaw*, *chain reaction* and *supermarket chain*, where the constituent *chain* has three different readings, depending on in which position it occurs in which compound (‘connected series of metal links’, ‘sequence, series’, or ‘a set of businesses controlled by one firm’, respectively). One way of addressing both issues is to use contextualised vectors, which supposedly reflect the reading instantiated in the respective textual context.

In the present study, we will use contextualised embeddings to represent compound meanings. The provenance and treatment of these embeddings will be discussed in detail in section 4.1.

In the next section, we will first look at how compound stress is realised in our sample, using pitch

contours as a correlate of stress. In section 4 we will then investigate the properties of the semantic vectors of the compound in our sample in relation to analyses of compound semantics that have used discrete semantic relations and semantic categories. In section 5 we will explain and implement the mapping of form and meaning in an LDL model.

## 3 Study 1: Compound phonetics and phonology

### 3.1 Methodology

#### 3.1.1 The data set

The data for this paper have been used in two previous empirical studies of compound stress, Plag et al. (2008) and Kunter (2011).<sup>2</sup> The data were made available to us by these authors. The number of tokens in this combined data set is N=4362.

To create the compound datasets for the original studies, a team of research assistants read the transcript of the corpus and identified all constructions consisting of exactly two nouns, regardless of whether they were transcribed as one orthographic word (e.g. *footsteps*) or two orthographic words (e.g. *foot patrol*). Subsequently, personal names such as *Thomas Finneran* and two-part appositive constructions such as *Governor Dukakis* were excluded; also excluded were types whose left constituent could be interpreted as an adjective, such as *administrative abilities*.

From 4,362 tokens in the data set we excluded 661 items because of background noise. In eight cases Praat was unable to extract pitch, three further items were excluded because they were part of a bigger compound (e.g. noun-noun-noun) and one was excluded because of mispronunciation. This resulted in a data set with N=3,689.

In order to take into account different contexts in which a given compound type occurs, we sampled only those types that occur in at least three different contexts, excluding duplicated contexts. Furthermore, to ensure that speaker variation also comes into play, we included only those types that were spoken by at least two speakers. As a result, each compound type comes with at least three unique contexts and with least two unique speakers for its tokens. Based on the previous data set of 3,689 tokens of 2,017 types, this more strictly controlled data set consists of 971 tokens of 150 types.<sup>3</sup>

---

<sup>2</sup>A number of subsequent studies have used subsets of the data, often with additional variables coded and analysed, e.g. Arndt-Lappe (2011); Plag & Kunter (2010); Plag (2010).

<sup>3</sup>The steep decline in the number of available data points reflects the underlying type-token distribution in the data. In the full dataset, 1,523 types occur only once and a further 255 types occur only twice. Such a heavy-tailed frequency distribution is typical of linguistic data and is often approximated by Zipf's law.

### 3.1.2 From raw pitch to pitch contours

To arrive at workable pitch data for the 971 compound tokens of interest, we implemented the following steps. First, individual pitch ranges were first determined for each speaker based on all available utterances using Praat (Boersma & Weenink, 2019). Second, speaker-specific pitch floors and ceilings based on the speaker-specific ranges were then applied during the extraction of the pitch data of the compound tokens. Third, voiced and unvoiced segments were identified using the Praat script by Al-Tamimi & Khattab (2015); Al-Tamimi (2018), and pitch values corresponding to unvoiced frames were replaced by NAs.<sup>4</sup>

Then, all contours were time-normalised to 51 equally spaced samples per token (2% intervals) to enable direct comparison across tokens of differing durations. Where there was no pitch data in the original data, this gap was retained in the time-normalisation.

In addition, we applied the same time-normalisation and sampling procedure separately to each constituent within the compound tokens. Constituent time axes were normalised to their own duration, and pitch values were extracted at 4% intervals, yielding 26 equally spaced time-normalised samples per constituent. To keep the pitch data of both time-normalisation approaches comparable, i.e. to arrive at 51 samples for each token, the 26th sample of the first constituent and the 1st sample of the second constituent were averaged. This within-constituent sampling revealed that 24 tokens came without pitch data in at least one of their two constituents and 2 tokens had only one pitch data point available in their first constituents. We dropped these 26 tokens from further analyses. Checking the data set selection criteria (cf. Section 3.1.1), we found that three types now no longer followed the sampling criteria, as they were represented by only two tokens, i.e. they no longer had at least tokens from three different contexts. Consequently, all tokens of the three types were excluded, resulting in a data set of 939 tokens.

The resulting pitch tracks, both for across and within constituent time-normalisation and sampling, contained micro-prosodic fluctuations and were missing intervals due to voiceless segments. In order to obtain smoothed, continuous pitch contours for all tokens, we used generalised additive models (GAMs; Wood, 2017). In using GAMs we follow other studies that investigated pitch contours in simplex words (Chuang et al., 2024; Jin et al., 2025) and compounds (Kösling et al., 2013). GAMs are particularly suited for this task because they are able to model time-varying patterns without assuming linearity and can infer likely values for voiceless portions based on data points of other tokens. In other words, the fitted GAMs provide statistically informed interpolations that fill in the pitch trajectory where direct measurement was not possible, while simultaneously smoothing out micro-variations that do not reflect meaningful prosodic modulation.<sup>5</sup>

---

<sup>4</sup>For full technical specifications, see Appendix A.

<sup>5</sup>In the present paper we did not make use of the PraatSmooth function used by Schmitz et al. (2026) but opted for smoothing

We fitted the GAMs using the *mgcv* package in R (R Core Team, 2019).<sup>6</sup> GAMs extend standard regression models by estimating smooth, non-linear functions for continuous predictors. These smoothing splines adapt flexibly to the data: they can follow gradual curvature in the contour but are penalised for unnecessary wiggleness, thereby striking a balance between fidelity to the data and generalisation to unseen values. In the present implementation, normalised time served as the main predictor, allowing us to model the full pitch trajectory across each token. We included a smooth for the effect of time by compound type (`s(timeNorm, type)`), which estimates a population-level contour for each compound type, capturing systematic temporal variation in pitch across tokens. A corresponding smooth for each speaker (`s(timeNorm, speaker)`) accounted for individual differences in baseline range and contour shape. These factor smooths act as non-linear random effects: rather than assigning a single intercept per speaker, they estimate a separate time-varying curve for each individual (Baayen et al., 2022).

Duration may also have an effect on pitch excursion, as tokens with longer durations give the speaker more time for larger excursions. To control for these effects, we added tensor-product smooths of time with the durations of the first and second constituent (`ti(timeNorm, durationN1)` and `ti(timeNorm, durationN2)`). Tensor-product smooths extend the idea of smoothing to interactions between predictors and here capture how the development of pitch over time is modulated by the duration of the constituents within compounds.

Pitch contours show massive autocorrelation as pitch changes gradually and continuously over time, such that it is very easy to predict the pitch value at a given point in time  $t + 1$  on the basis of the pitch value at point in time  $t$ . Autocorrelation leads to autocorrelated residuals in regression models, which violates the central assumption of independence of residuals (Baayen et al., 2018b, 2022). To address the autocorrelation problem we followed established procedures (see, for example, Baayen et al. 2022; Chuang et al. 2024) and incorporated a first order autoregression model ('AR(1)') into our GAM models. This is a linear model that predicts pitch on the basis of the preceding value. We determined the appropriate value of  $\rho$  in AR(1) by first fitting a GAM without AR(1), and then calculated the autocorrelation at lag 1, then fitted the GAM again with AR(1) with  $\rho$  set to the lag 1 autocorrelation.

The resulting GAM was then used to generate predicted pitch contours for each of the 971 compound tokens. For every token, the fitted GAM produced a complete pitch contour across the entire normalised time course, including interpolated values for missing segments. These model-based contours thus represent statistically informed estimates of the underlying, continuous pitch movements associated with each com-

---

through GAMs for the reasons just given. In contrast to GAM-smoothing, the `PraatSmooth` function has no information about the type, smoothes linearly (introducing unnatural pitch plateaus), and cannot be corrected for autocorrelation, which is a massive problem for any smoothing algorithm (see below for discussion). In sum, one can expect that the data derived through the `PraatSmooth` function are much more noisy and less trustworthy than those derived through GAM modelling.

<sup>6</sup>All data sets and scripts are available at [https://osf.io/gdyjr/overview?view\\_only=ac6afcc674ab4b6d8e24daabee4d67f](https://osf.io/gdyjr/overview?view_only=ac6afcc674ab4b6d8e24daabee4d67f).

pound token, independent of local micro-prosodic noise. Figure 3 illustrates this process for the compound *health care* in a reduced example data set consisting of ten tokens of the compound type *health care*, which were predicted in a GAM with the specifications just outlined.

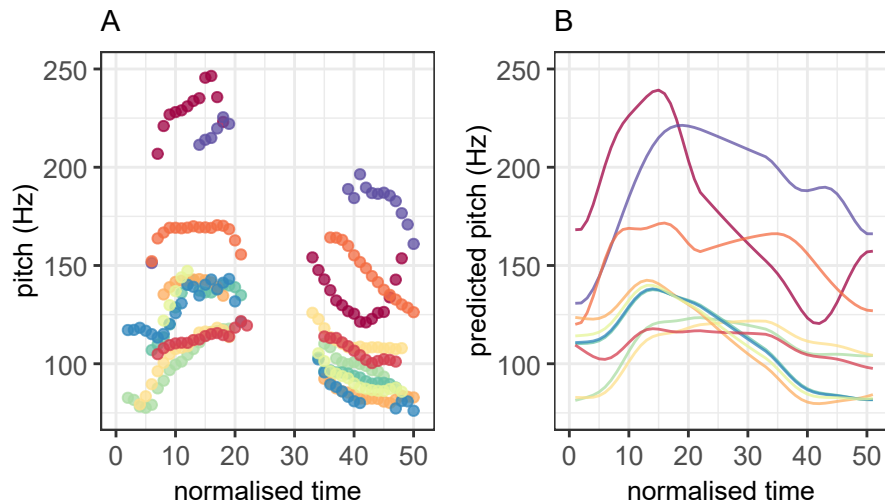


Figure 3: Example data. Panel A shows the pitch points of single tokens of the compound type *health care* produced by different speakers in different contexts. Panel B shows the individual pitch contours as predicted by the GAM.

The pitch data with gaps for voiceless stretches is illustrated in panel A of Figure 3. Between time steps 20 and 35, there are no pitch data available, corresponding to the voiceless dental fricative at the offset of *health* and the voiceless velar plosive at the onset of *care*. Similarly, none of the contours start at time step 0, as there is no pitch data available due to the voiceless glottal fricative at the onset of *health*. Across the ten example tokens, we see considerable variation in overall pitch height, in the precise timing of the voiceless regions, and in the extent of micro-prosodic perturbations; patterns likely driven by both inter- and intra-speaker differences.

Panel B of Figure 3 shows the corresponding fitted contours produced by the GAM. These curves represent the model’s best estimate of the continuous F0 trajectory for each token, given the information available: token identity, compound type, speaker identity, and the durations of the first and second constituent. The fitted curves broadly follow the observed data where measurements exist, while providing smooth, data-informed interpolations for intervals in which the signal was unvoiced.

In order to investigate the relation of the pitch contours to putative more abstract categories of compound prominence, we implemented k-means clustering for longitudinal and trajectory data using the *kml* package in R (Genolini et al., 2015), treating each compound token’s pitch contour (i.e. its 51 time-normalised mean-centred pitch values in semitones) as a time series. This method clusters trajectories based on their overall

shape, grouping tokens with similar pitch movement patterns across time. To avoid poor local optima, the algorithm was repeated 20 times for each cluster solution, and the best result was retained. The algorithm begins by selecting an initial set of cluster centroids, provisional representative contours, usually chosen at random from the data. Each individual trajectory is then assigned to the cluster whose centroid is closest to it, measured by overall Euclidean distance across the 51 points. The centroid of each cluster is then recomputed as the average contour of all trajectories assigned to it, and the assignment step is repeated. This iterative process continues until the centroids no longer change substantially, meaning the solution has stabilised. In other words, the algorithm groups together contours that rise and fall in a similar way, and it represents each group by an average contour that summarizes the general pattern of that group.

As input for the k-means clustering algorithm we used the pitch values predicted by the GAMs (cf. panel B) of Figure 3) corrected for inter-token and especially inter-speaker differences via centring and scaling, so that the clustering algorithm focusses on the location and amplitude of pitch movements rather than their overall height. For each token, we calculated the mean and range of its 51 predicted pitch values, subtracted the mean from each value, and divided by the range.

To determine the optimal number of clusters, we inspected the automatically generated summary for 2 to 15 clusters and selected the solution with the highest Calinski-Harabasz index (Caliński & Harabasz, 1974), which favours solutions with high between-cluster variance and low within-cluster variance.

### 3.2 Results: Pitch contours

Using k-means clustering as described above, the analysis yielded the 2-cluster solution as the best fit. This suggests that the pitch contours in our data are best described by two distinct trajectory types. The pitch contours grouped by the cluster they were assigned to are illustrated in panel A of Figure 4.

Cluster 1 resembles what can be described as left-stressed, i.e. a rising and falling contour making up a pitch accent in the first half of the compound. Cluster 2 appears to show more variation between the individual contours. However, overall its average smooth seems to represent a rise in the first part of the compound towards a clear peak and fall in the second half of the compound.

To ensure that the clustering as well as the manual inspection of the clusters is not affected by the timing of the individual constituents across different compound types and tokens, we conducted a second analysis, which takes the two constituents into account. The underlying data for this analysis are the compound pitch contours predicted by the GAM based on the within-constituent time-normalised and sampled data.

We then applied the same k-means clustering procedure to the within-constituent time-normalised pitch data. Again, a 2-cluster solution yielded the best fit according to the Calinski-Harabasz index. The clusters

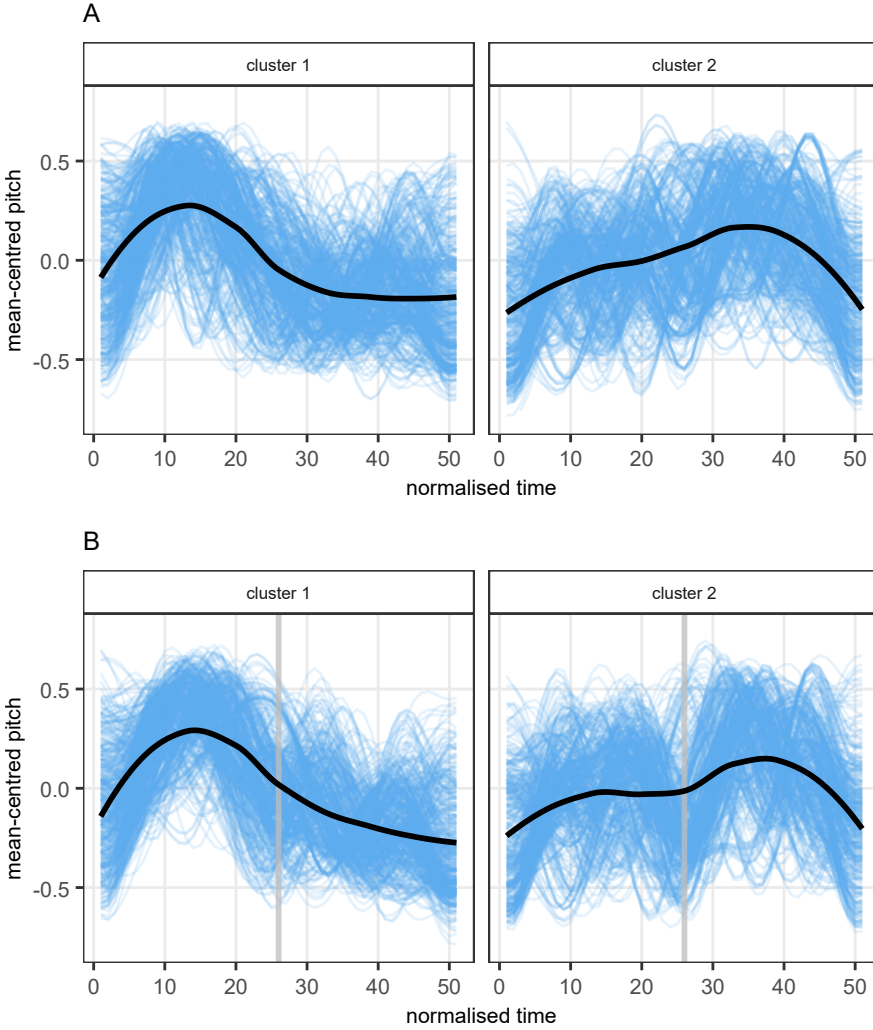


Figure 4: Pitch contours of the two clusters, with a non-linear average smooth indicating the general trend in the data. Panel A across-constituent time-normalised data; panel B within-constituent time-normalised data. The vertical line in panel B is at the constituent boundary.

identified in this analysis are illustrated in panel B of Figure 4. Cluster 1 looks as before, while cluster 2 appears to be a bit more structured and interpretable. That is, in the first constituent there is a small rise, followed by slight fall or plateau. In the second constituent there is a pronounced rise followed by a fall, which probably reflects stress placement within the constituent.

In both cluster analyses, compound tokens in cluster 1 show an early peak, which corresponds to prominence on the first constituent. In cluster 2, on the other hand, the compound tokens show a late peak, which can be interpreted as indicating prominence on the second constituent.

The interpretation of cluster 2 is hampered by the fact that the overlay of the curves makes it hard to discern and assess individual curves, more so than in cluster 1. In order to address this problem we have

plotted fifteen random samples from cluster 1 of three compound tokens each in Figure 5.

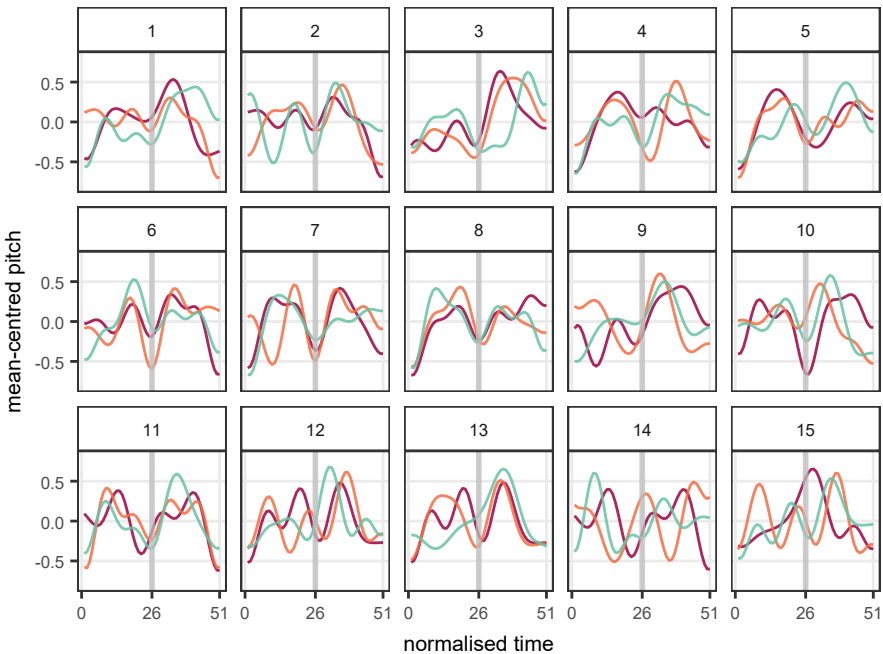


Figure 5: Pitch contours of fifteen samples of three tokens each from cluster 2. Each line represents one token, with within-constituent time-normalised data. The vertical line in each panel is at the constituent boundary.

We can now see that most of the compounds sampled from cluster 2 have positive pitch excursions in both the first and the second constituent, which can be interpreted as the presence of two pitch accents. Notably, the excursions are not as pronounced as with the left-stressed compounds, which is in line with what previous phonetic studies of compound stress or primary vs secondary stress have found (see again Kunter & Plag 2007; Kunter 2011; Plag et al. 2011). Overall, the plots in Figures 4 and 5 seem to lend themselves to an interpretation along the lines of the traditional binary categorisation of compound prominence.

Further inspection also allows us to look into the problem of stress variation within a given type. In the across-constituent time-normalised data, 52.8% ( $n = 513$ ) of all tokens fall into cluster 1. At the type level, 126 of 150 types are assigned to one of the two clusters with at least 80% of their tokens; 110 types have all their tokens in a single cluster (1: 70; 2: 40). Among the 73 most frequent types, i.e. those with five or more tokens, 64 types find at least 80% of their tokens in one cluster; 48 types have all their tokens in a single cluster (1: 26; 2: 22).

In the within-constituent time-normalised data, we find a similar picture. 52.6% ( $n = 494$ ) of all tokens fall into cluster 1. At the type level, 114 of 147 types are assigned to one of the two clusters with at least 80% of their tokens; 96 types have all their tokens in a single cluster (1: 57; 2: 39).

We will now turn to the question of how these pitch contours relate to the semantics of the compounds at hand.

## 4 Study 2: Compound semantics

### 4.1 Methodology

It is currently not quite clear what the best way is to obtain semantic vectors (‘embeddings’) for compounds, especially if they are spelled as two words. Very simple models are the additive and multiplicative models (Mitchell & Lapata, 2010), which, however, have the serious disadvantage of not taking the order of the constituents into account, due to commutability. Alternatively, one could use whole-word orthographic representations of compounds. This choice would necessitate first a step of artificially creating closed forms, even if the compound is originally written as two orthographic words. Based on these modified orthographic representations one can derive vectors using established distributional semantic models (e.g. word2vec, Mikolov et al., 2013). One problem with this approach is that most compounds are of very low frequency, with the consequence that reliable vectors are hard to obtain in this way from off-the-shelf semantic spaces. One might also think of using the CAOSS model (Günther & Marelli, 2021), which implements position-specific matrices and matrix multiplication to derive compound vectors from their constituent vectors. All these approaches have the disadvantage that they cannot be used straightforwardly to obtain contextualised vectors.

To obtain contextualised semantic embeddings for all 3689 compound tokens, we followed another procedure (see, for example, Chuang et al. 2024; Jin et al. 2025, for analogous work) and used the pre-trained BERT model *bert-base-uncased*, which is based on 3.3 billion words with 110 million parameters.<sup>7</sup> The aim was to compute vector representations that reflect the meaning of a word as shaped by its immediate linguistic surroundings. Code was implemented in Python, using the *transformers* library (Wolf et al., 2020), which provides access to pre-trained transformer models, their associated tokenisers, and tools to handle them.

For each compound token, the input for BERT was composed of not only the compound but also its preceding and following sentence context. This input was tokenised using BERT’s tokeniser, which segments the input into subword units. This step is required as BERT does not operate on whole words but instead processes sequences of subword tokens, allowing it to flexibly handle morphologically complex or previously unseen words.

---

<sup>7</sup>The model is available at [https://huggingface.co/docs/transformers/en/model\\_doc/bert](https://huggingface.co/docs/transformers/en/model_doc/bert).

To ensure accurate alignment between the target compound and its position in the model’s tokenised input, the target compound was also tokenised independently, and its subword sequence was then located within the tokenised context. The full input was passed through BERT, which returns a sequence of so-called hidden states, i.e. high-dimensional vectors generated by each layer of the model’s network as it processes the input. These vectors are called “hidden” because they represent internal computations of the model that are not directly observed. Specifically, we used the final-layer hidden states, which correspond to the model’s most refined representation of each token after integrating information from the entire input sequence.

Each hidden state has a fixed dimensionality of 768 and encodes both the lexical identity of a token and the contextual relationships it has with the other tokens in the input. For example, the hidden state for the word *bank* would differ depending on whether the context refers to a financial institution or a riverbank, because the model incorporates surrounding words into the representation.

The hidden states corresponding to the subword tokens of the target compound were extracted and averaged, resulting in a single vector representation per target compound. This averaging procedure ensures that words split across multiple subword tokens are represented as a unified embedding. It was chosen as a straightforward and interpretable method for aggregating subword information, preserving the contribution of each part of the words that make up the compound.

This embedding procedure was applied to all compound tokens and their contexts in the dataset, yielding context-sensitive vector representations for each compound token.

## 4.2 Results: Semantics

**Types and tokens.** To explore the semantics of the present set of compounds, we first used t-distributed Stochastic Neighbour Embedding (t-SNE; van der Maaten & Hinton 2008) to project the high-dimensional BERT-based embeddings into a two-dimensional space. The perplexity parameter, which controls the balance between local and global structure preservation, was optimised through visual inspection. We generated t-SNE plots using a range of perplexity values (5, 10, 20, 30, 40, 50) and compared the resulting plots in terms of cluster coherence and interpretability. A perplexity value of 20 produced the clearest separation of type-related compound tokens. This setting was therefore chosen for the final visualisation of the compound semantics, as shown in Figure 6. Overall, tokens of a compound type form rather clear clusters, in which non-identical t-SNE dimensional values reflect different contexts.

**Semantic relations and categories.** As described in Section 2.2, various semantic relations and semantic categories have been shown to correlate with different prominence patterns. As, in this paper, we are using experience-based semantic vectors instead of distinct semantic relations and categories, it is an interesting

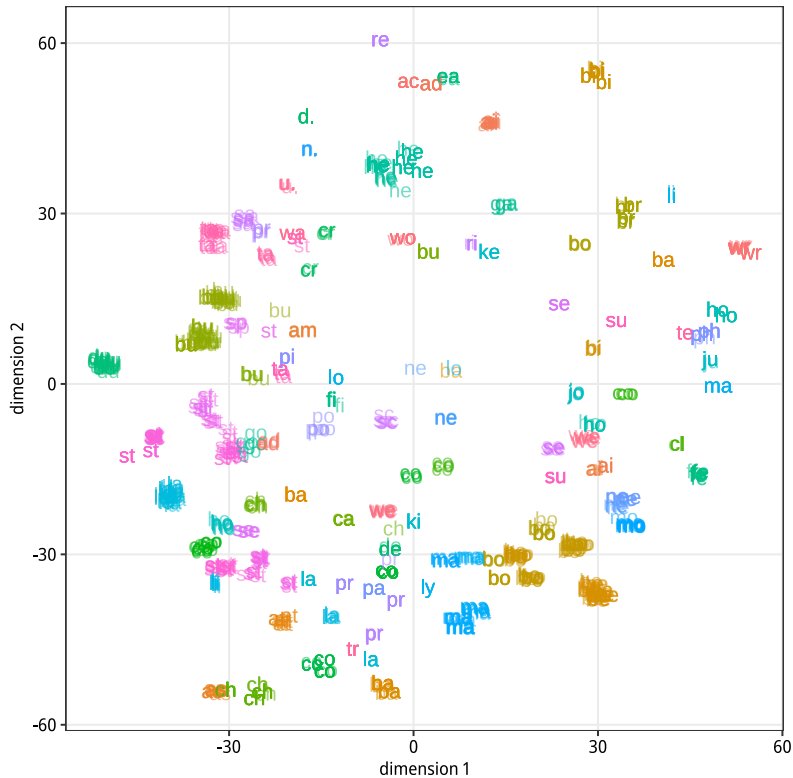


Figure 6: t-SNE plot of compound embeddings of types with at least five tokens in the BERT semantic space. Tokens of the same type match in color and letter strings.

question whether such vectors contain information about these semantic properties of the compounds in our sample.

In a study of the relationship between semantic vector representations and relational properties of compounds, Günther & Marelli (2022) demonstrated that semantic vectors can be used to predict relational interpretations by human participants in an experiment well above chance level. This means that relational information (which is usually hand-coded using categories like ‘N2 is made of N1’, ‘N2 is located at N1’) is indeed encapsulated in distributional vectors.

In order to test whether our embeddings also bear out information about the semantic relation between compound constituents and categorically coded semantic properties of constituents and compounds, we used the ratings that came with the data set that we are using in the present study, and which have been used already in previous studies of semantic effects on compound stress (e.g. Plag et al., 2008). Each compound was coded independently by two raters, who did not interpret and categorise the compounds in isolation (i.e. as types), but took into account the context in which the compounds occurred in the news texts. Hence, the coding was token-based. Due to the well-known problems of constituent ambiguity, compound ambiguity,

and the vagueness of the definitions of the semantic relations and categories, a compound may be rated to exhibit more than one relation or category. As a result, each compound token in the data set has two or more ratings, with one or more by each rater. In the reduced data set of 971 tokens that we use in this paper, each compound token has on average 2.7 (instead of the minimum of two) ratings. The ratings of the two raters often coincide, but often do not. For example, in the original study by Plag et al. (2008), 2,041 of the 4,353 tokens received the same rating of some semantic category or relation.

In order to investigate whether the embeddings include relational information or information about the semantic categories, we devised a multivariate multiple linear regression model to predict the ratings for the 18 semantic relations and eight semantic categories coded for each compound token from the 768 dimensional values of its contextual semantic vector. This regression model differs from more common regression models in that more than one outcome variable is predicted at the same time. To minimize the risk of overfitting, we employed cross-validation using the `cv.glmnet` function from the `glmnet` package (Friedman et al., 2010; Simon et al., 2011; Tay et al., 2023). The model estimates a regression weight between each predictor and each outcome variable. We chose the highest weight as the most likely semantic value, and then compared whether the semantic value predicted by the regression model is the same as one of the values chosen by the raters. If that was the case, the prediction was counted as accurate. We devised four models, two for the whole data set (N=3,689), and two for the reduced data set (N=971). For each data set we devised one model using the 768 dimensions, and one with a reduced number of dimensions, thereby addressing potential concerns of overfitting. The dimensions were reduced with the help of a principal components analysis (using the `prcomp` function in R). We kept the first principal components that together explained at least 80% of the variance. This resulted in a reduction to 158 dimensions for the full data set, and 88 dimensions in the reduced data set.

We first tested the whole data set. The model using 768 dimensions predicts 56% of the ratings correctly (2,065 out of 3,689 ratings), while the model with only 158 dimensions still predicts 47% correctly. The baseline accuracy is only 3.8% (one out of 26). This means that the vectors indeed incorporate information about the semantic properties that have (or have not) a say in determining the stress pattern of compounds.

The reduced data set has only 150 types, as against 2,017 types in the full data set. Given that the contextualised tokens of one type will be very similar to each other (as shown in Figure 6 above), the regression model has very limited information (i.e. 150 types) to predict 26 different outcomes. Hence, we can expect that the semantic vectors of the reduced data set will not be as successful in predicting semantic relations and categories as the contextualised vectors of the 2,017 types in the full data set. This expectation is borne out by the analysis. Testing the reduced data set yields much lower accuracy rates. Both models (based on either 768 or 88 dimensions) have an accuracy of 11%. This accuracy is larger than the baseline,

but not significantly so ( $p = 0.35$ , Fisher’s test).

**Semantics and contour clusters.** Given that some semantic relations and categories have a tendency to go together with particular stress patterns, and that this semantic information is also incorporated in the embeddings, one might hypothesise that it is possible – at least to some extent – to predict a compound’s stress pattern directly from the semantics. We tested this hypothesis by implementing a logistic regression analysis (using the `glm` function in R) in which we used the dimensions of the semantic vectors to predict whether a given compound belongs to cluster 1 or to cluster 2.

We first fitted a model to the 971 compound tokens with all semantic dimensions as predictors, expecting over-fitting. This was indeed the case, the concordance index for this model was 1.0 after rounding after the second digit. We then fitted a second model with the 88 principal components introduced above. This model yielded a concordance index of 0.77, indicating little danger of overfitting and rather moderate predictive power. A third model using 50 dimensions yielded a concordance index of 0.75.

The latter two models can be interpreted in such a way that it seems to be possible, with moderate success, to predict the stress pattern of compounds from their semantic vectors. In the following section we will further explore the relation between meaning and prosody in compounds by implementing a discriminative learning model.

## 5 Study 3: Mapping prosody and semantics

### 5.1 Methodology

The present study uses a particular implementation of a learning theory, Linear Discriminative Learning (LDL, e.g. Baayen et al. 2018a, 2019), to explore the role of semantics in compound prosody.

Discriminative learning theory is a well-established theory of learning from cognitive psychology (e.g. Rescorla 1988a; Pearce & Bouton 2001). The general cognitive mechanisms assumed in this theory have been shown to be able to model a number of important effects observed in both animal and human learning, for example the blocking effect (Kamin 1969) and the feature-label ordering effect (Ramsar et al. 2010).

Discriminative learning has recently been introduced to linguistics, and numerous studies have shown that it can successfully model important problems in phonetics, phonology and morphology (see Plag 2018; Lieber 2021 for linguistic textbook introductions). The central assumption of discriminative learning theory is that learning results from exposure to informative relations among events in the environment. These relations, or ‘associations’, can then be used to build representations of the world around us. The associations (and the resulting representations) are constantly updated based on new, informative experiences. The events

associated with each other are called ‘cues’ and ‘outcomes’, and the association between cues and outcomes is computed mathematically using the so-called Rescorla-Wagner equations (Rescorla & Wagner 1972; Rescorla 1988a,b). The equations work in such a way that the association strength or ‘weight’ of an association increases with every time that a given cue and a given outcome co-occur. Conversely, the weight decreases whenever a given cue occurs without that outcome. Towards the end of the learning process, a stable final state is asymptotically approached, with final association weights. The final association weights can be conceived as the activation of particular outcomes based on the training with all cues. This type of model is known as ‘Naive Discriminative Learning’ (NDL).

Based on the tenets of discriminative learning, Baayen and colleagues have developed a theory of the mental lexicon, called the ‘Discriminative Lexicon’ (see Baayen et al. 2019; Chuang & Baayen 2021). This theory implements a computational architecture which grew out of NDL, and is called ‘linear discriminative learning’. LDL generates a system of form-meaning relations by discriminating between different forms and meanings (instead of building them from compositional units, as in morpheme-based morphology, for example). In an LDL approach, forms are represented by numerical vectors, and meanings are also represented by numerical vectors. The idea is that, if both forms and meanings can be expressed numerically, we can mathematically connect the two levels of representation, i.e. map meaning onto form, or form onto meaning.

In this system of learning, the two sets of vectors are combined into matrices – a form matrix and a meaning matrix. The form vectors are mapped onto meaning vectors to model comprehension, and meaning vectors are mapped onto form vectors to model production. The mapping between them at the theoretical end-state of learning is estimated using multivariate multiple linear regression (hence the term ‘linear discriminative learning’). The network is simple and interpretable, because, in contrast to deep learning networks, it features just two layers (i.e. the form and meaning matrices), both of which are linguistically transparent.

There are two ways of testing and using these networks. One possibility (‘internal validation’), is to have the model generate word forms or meanings that can then be compared to empirically observed word forms or meanings (e.g. Baayen et al. 2018a, 2019; Van de Vijver & Uwambayinema 2022; Nieder et al. 2022). The other possibility of using the networks (‘external validation’) is to derive secondary measures from the associations given by the networks (Heitmeier, 2022), and to use these measures to predict things outside the network, e.g. independent properties of words, or human behaviour (e.g. Stein & Plag 2021; Schmitz et al. 2021; Plag et al. 2025).

In what follows we will do both, starting with the accuracy of predicting the correct pitch contour or the correct semantics.

### 5.1.1 The LDL model

To train an LDL network, two matrices are needed, one of them representing the meaning of the compounds, and the other one representing their form. Figures 7 and 8 give a toy example each of a form matrix and a semantic matrix, respectively. In an LDL implementation, these matrices are mapped onto each other using the matrix algebra of multivariate multiple regression (see, for example, Chuang & Baayen, 2021, in which the mathematical underpinnings of LDL implementations are described in detail). In the present paper, pitch contours (i.e. sequences of pitch values) are used as cues (illustrated as `cue1` through `cue5` in Figures 7, 9 and 10), and word embeddings as semantic vectors (illustrated with the dimensions `S1` through `S4` in Figures 8, 9 and 10).

$$\mathbf{C} = \begin{array}{cc} & \begin{array}{ccccc} \text{cue1} & \text{cue2} & \text{cue3} & \text{cue4} & \text{cue5} \end{array} \\ \begin{array}{c} \text{chief justice} \\ \text{retirement age} \end{array} & \left( \begin{array}{ccccc} 1.07 & 1.96 & 2.79 & 3.51 & 3.69 \\ 2.13 & 1.91 & 1.79 & 1.41 & 1.12 \end{array} \right) \end{array}$$

Figure 7: Toy  $\mathbf{C}$  matrix with acoustic cues.

$$\mathbf{S} = \begin{array}{cc} & \begin{array}{cccc} \text{S1} & \text{S2} & \text{S3} & \text{S4} \end{array} \\ \begin{array}{c} \text{chief justice} \\ \text{retirement age} \end{array} & \left( \begin{array}{cccc} 0.69 & 0.18 & 0.73 & 0.61 \\ 0.50 & 0.11 & 0.09 & 0.81 \end{array} \right) \end{array}$$

Figure 8: Toy  $\mathbf{S}$  matrix with semantic vectors

As we are interested in whether a compound’s meaning can be predicted from its form and whether a compound’s form can be predicted from its meaning, we implemented LDL in a leave-one-out approach. In this approach, the transformation matrices  $\mathbf{F}$  and  $\mathbf{G}$  (as illustrated in Figures 9 and 10) are trained on all but one compound token. Using these versions of  $\mathbf{F}$  and  $\mathbf{G}$ , the semantics of the left-out compound token is predicted for comprehension and the form of the left-out compound token is predicted for production. In other words, the knowledge of the discriminative lexicon as reflected in  $\mathbf{F}$  and  $\mathbf{G}$  is used to comprehend and produce a token novel to the respective LDL network. This process was applied to all tokens of the data set, i.e. each token was left-out once.

$$\mathbf{F} = \begin{matrix} & \mathbf{S1} & \mathbf{S2} & \mathbf{S3} & \mathbf{S4} \\ \mathbf{C1} & \left( \begin{array}{cccc} 1.8033 & 0.4269 & 0.9728 & 2.3780 \\ 2.3074 & 0.5629 & 1.6027 & 2.7427 \\ 2.8201 & 0.6991 & 2.1978 & 3.1518 \\ 3.1269 & 0.7869 & 2.6892 & 3.2832 \\ 3.1061 & 0.7874 & 2.7945 & 3.1581 \end{array} \right) \end{matrix}$$

Figure 9: Toy  $\mathbf{F}$  matrix

$$\mathbf{G} = \begin{matrix} & \mathbf{C1} & \mathbf{C2} & \mathbf{C3} & \mathbf{C4} & \mathbf{C5} \\ \mathbf{S1} & \left( \begin{array}{ccccc} 1.8033 & 2.3074 & 2.8201 & 3.1269 & 3.1061 \\ 0.4269 & 0.5629 & 0.6991 & 0.7869 & 0.7874 \\ 0.9728 & 1.6027 & 2.1978 & 2.6892 & 2.7945 \\ 2.3780 & 2.7427 & 3.1518 & 3.2832 & 3.1581 \end{array} \right) \\ \mathbf{S2} & \\ \mathbf{S3} & \\ \mathbf{S4} & \end{matrix}$$

Figure 10: Toy  $\mathbf{G}$  matrix

### 5.1.2 Representing meaning

To represent the meanings of the compound tokens in the LDL framework, we used the contextualised semantic embeddings introduced in Section 4. These embeddings provide token-specific semantic vectors that reflect how the meaning of a compound is shaped by its immediate linguistic context, rather than treating compound types as fixed points in semantic space. In contrast to static distributional models, which cannot readily distinguish between multiple senses or contextual nuances, contextualised embeddings capture fine-grained variation in usage by integrating information from the surrounding context. For the present study, each of the 971 compound tokens is therefore represented by a 768-dimensional vector.

### 5.1.3 Representing form

For the LDL production model, each compound token requires a form vector with a fixed dimensionality. Because the raw pitch tracks vary in duration and contain gaps due to voicelessness, the raw pitch values cannot be used directly for this purpose. The GAM-based contours introduced in Section 3 provide exactly the kind of representation needed here: they offer smooth, continuous trajectories based on 51 equally spaced time points per token. These fitted contours retain the characteristic shape of the pitch movement while supplying interpolated values where no pitch was available, and thereby provide a uniform basis for the LDL form space.

However, unlike the semantic embeddings, the predicted pitch contours bear the imprint of speaker-

specific differences in baseline pitch, pitch range, and habitual prosodic style. To place the form and meaning spaces on a comparable footing, and to prevent tokens from speakers with larger pitch ranges from disproportionately influencing the mapping, we applied token-wise centring and scaling to the predicted contours (cf. Chuang et al., 2024). For each token, we calculated the mean and range of its 51 predicted pitch values, subtracted the mean from each value, and divided by the range. This min-max normalisation removes speaker-dependent amplitude and baseline differences.

## 5.2 Results: Mapping pitch and semantics

### 5.2.1 Finding the right form, finding the right semantics

The accuracy of modelling comprehension and production was checked via an analysis of nearest neighbours. If the predicted vector of a given compound token was most similar to the input vector of another token of the same type, the prediction was considered to be accurate. In other words, if the predicted semantics or form of a token of a given compound type was most similar to a token of the same type, we assume that this token was correctly comprehended or produced, respectively. The results of this analysis for the pitch data time-normalised across and within constituents are given in Table 1. We compare the model accuracies to a baseline probability, which was calculated by determining the relative frequency of each type across all tokens and averaging these values, yielding an estimate of the expected likelihood of encountering a given type.

Table 1: Accuracy via nearest neighbours for comprehension and production in the LDL implementations based across-constituent and within-constituent time-normalised pitch data.

<b>time-normalisation</b>	<b>comprehension</b>	<b>production</b>	<b>baseline</b>
across constituents	19.6	13.3	1.17
within constituents	20.2	12.5	1.19

We can see that irrespective of the time-normalisation approach we reach an accuracy of about 20 percent for comprehension and about 13 percent in production. This is a vast improvement over the baseline accuracies of only 1.17 and 1.19 percent.

To better understand why comprehension outperforms production, we examined the structural properties of the semantic and pitch-based spaces in a post-hoc analysis.<sup>8</sup> In particular, we were interested in understanding how easy (or hard) it is to detect a certain type in comprehension vs. production by looking at the variability between the tokens of a type, and between different types, both on the form side and on the semantic side. For instance, if tokens of a type are more similar to each other, and types differ more

<sup>8</sup>The post-hoc analysis was carried out for both across- and within-constituent time-normalised pitch spaces and the corresponding semantic matrices. As the resulting distributions were effectively identical, we do not distinguish between the two in what follows.

clearly from other types in one of the two spaces, it is easier to detect the correct type in that space than in the other space.

We quantified how tightly tokens cluster within compound types (intra-type similarity) and how distinct different types are from one another (inter-type similarity) in both the semantic matrix and the form matrix. Because the semantic vectors have 768 dimensions whereas the pitch vectors have 51, we first reduced both representations to a shared dimensionality using principal component analysis, retaining the first 51 components in each case. Working with this reduced space ensures that cosine-based similarity measures are directly comparable across both matrices.

Intra-type similarity was computed by taking, for each compound type, all token vectors belonging to that type and calculating the mean of all pairwise cosine similarities among them. This measure captures how diverse the vectors for a given compound type are. Inter-type similarity was assessed by computing a centroid for each type and then calculating, for each centroid, the average cosine similarity between it and the centroids of all other compound types. This yields a measure of how distinct each type is relative to the rest of the lexicon. These steps were carried out for the semantic space and for the pitch space, producing directly comparable distributions of intra- and inter-type similarities.

The resulting patterns show that semantic tokens form slightly tighter clusters within types and exhibit lower average similarity across types than pitch tokens. In other words, types are more internally coherent and more clearly separated in the semantic space than in the pitch space. These observations are confirmed by Wilcoxon tests. Figure 11 visualises these distributions for both intra- and inter-type similarity, showing the denser semantic clustering and the broader pitch distributions.

These structural properties provide a basis for understanding the asymmetry between comprehension and production in the LDL models. At first glance, the dimensionality of the representational spaces might suggest that production should be easier, as it maps from a 768-dimensional semantic vector to a 51-dimensional pitch vector, whereas comprehension maps from the lower-dimensional pitch vector to the higher-dimensional semantic vector. However, the difficulty of a linear mapping is not determined by the number of output dimensions alone; it depends on how much systematic, task-relevant structure is present in the input space.

In comprehension, the model receives pitch-based form vectors as input. Although these vectors have fewer dimensions, they contain systematic cues that reliably reflect the pitch realisation of the compound types in our data set. These cues are sufficient for the model to identify the appropriate region of semantic space for a given token. At the same time, our intra- and inter-type analyses show that the pitch space is globally less structured than the semantic space: tokens belonging to the same compound type are more dispersed, and the type centroids are less clearly separated. Despite this weaker global structure, the pitch

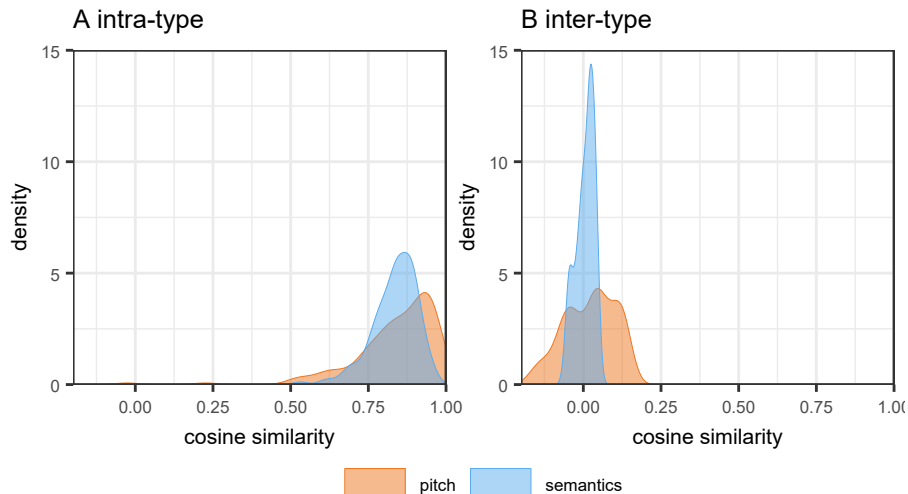


Figure 11: Distributions of cosine similarities in the semantic space and the pitch space. Pitch contours were those predicted by the informed GAM with across-constituent time-normalisation. Panel A shows intra-type similarity, computed as the mean pairwise cosine similarity of all tokens belonging to the same compound type. Panel B shows inter-type similarity, computed as the mean cosine similarity between each type centroid and the centroids of all other types.

vectors contain enough local regularity for the mapping from pitch to semantics to succeed.

In production, the model must map from the richly structured semantic space into the lower-dimensional pitch space. Here the difficulty is in the inverse of comprehension. The semantic vectors encode many distinctions that have no systematic influence on the acoustics, and only part of the semantic variation corresponds to consistent differences in pitch contours. The mapping therefore has to compress high-dimensional semantic structure into an output space that is comparatively noisy and exhibits weaker overall separation between types. This compression problem, combined with the limited discriminability of the pitch space, makes accurate production predictions substantially more difficult.

Taken together, this post-hoc analysis shows that the asymmetry between comprehension and production is not a matter of output dimensionality but of input structure and target discriminability. Comprehension benefits from pitch cues that are locally systematic and from a semantic space that is both coherent within types and well separated across types. Production, in contrast, requires the mapping of the well-structured semantic vectors onto a much less structured pitch space, which limits the achievable accuracy.

### 5.2.2 Inspecting other properties of the model

In order to further inspect some of the properties of the LDL model we extracted a range of production and comprehension measures from the network. The comprehension measures are semantic in nature and quantify semantic properties of the target word itself, or its relation to other words as predicted from the

form matrix. The production measures are phonetic in nature and – analogously to the semantic measures – quantify properties of the target’s form, i.e. its pitch contour, or of its relation to other forms as predicted from the semantic vectors. The general idea behind the inspection of these measures is to see whether the lexical properties predicted by and represented in the model can be related to the stress pattern a given token shows. In other words, we want to ask whether we can predict a token’s membership in cluster 1 or 2 on the basis of the comprehension and production measures extracted from the LDL model.

The measures are listed in Table 2 and they will be explained and discussed in detail below. We used measures that have been successfully used in previous work to investigate lexical, morphological, or morpho-phonological problems (e.g. Stein & Plag 2021; Schmitz et al. 2021, 2023; Plag et al. 2025), but we also computed new measures.

Table 2: LDL measures

LDL measure	Description
L1NORM	City block distance of $\hat{\mathbf{s}}$ or of $\hat{\mathbf{c}}$
L2NORM	Euclidean distance of $\hat{\mathbf{s}}$ or of $\hat{\mathbf{c}}$
DENSITY	Density (correlation-based)
ALC	Average Lexical Correlation
EDNN	Euclidean Distance to Nearest Neighbour
NNC	Nearest Neighbour Correlation

L1NORM and L2NORM: Both measures compute the length of the predicted vector  $\hat{\mathbf{s}}$  or  $\hat{\mathbf{c}}$  of a target form. The L1NORM is the sum of the absolute values of vector elements of a given word’s predicted vector, i.e. its city-block distance. In contrast, the L2NORM is computed as the square root of the sum of the squared values of  $\hat{\mathbf{s}}$ , i.e. its Euclidean distance. For the semantics, vector length can be conceptualised as a measure of semantic specificity. Vectors become shorter with increasing number of contexts in which a word appears, i.e. with decreasing semantic specificity (cf. Schakel & Wilson, 2015, 4). Longer vectors indicate more specific semantics.<sup>9</sup> For form vectors, a higher value of L1NORM or L2NORM is an indication of a more variable pitch contour, i.e. an indication of more pronounced excursions and/or of more oscillations in the pitch contours.

DENSITY: Density is a measure of the relationships (semantic or formal) of a target word with other words. This measure is derived for semantics by computing the mean correlation of the target’s  $\hat{\mathbf{s}}$  with the semantic vectors of its top eight neighbours in  $\mathbf{S}$  in terms of Pearson correlation. For the form, it is, analogously, the mean correlation of the target’s  $\hat{\mathbf{c}}$  with the pitch vectors of its top eight neighbours in  $\mathbf{C}$ . The higher DENSITY, the more similar the target word is to its neighbours, indicating a denser neighbourhood.

<sup>9</sup>It is thus to be expected that vector length and lemma frequency are negatively correlated (more frequent words tend to be less specific in meaning). This has been tested and confirmed in some studies. For instance, Plag et al. (2025) found the following correlations between vector length and log-transformed lemma frequency for German nouns: L1Norm:  $\rho = -0.25$ ,  $p < 0.0001$ , L2Norm:  $\rho = -0.26$ ,  $p < 0.0001$ , Spearman test.

ALC: Being another measure of lexical relatedness, Average Lexical Correlation is the mean value of all correlation values of a target’s estimated vector  $\hat{\mathbf{s}}$  or  $\hat{\mathbf{c}}$  with all vectors in  $\mathbf{S}$  or  $\mathbf{C}$ , respectively. Higher ALC values indicate a stronger connection of the target word’s meaning (or form) to the words’ meanings (or forms) in the lexicon at large.

EDNN: This variable focuses on the semantic or formal relationship of a target word with the word that is closest to it. The EDNN measure gauges the Euclidean Distance between a target’s predicted vector and its nearest neighbour’s vector in  $\mathbf{S}$  (or  $\mathbf{C}$ ) in terms of Euclidean distance. A lower value (i.e. smaller distance) indicates that this target’s nearest neighbour is semantically very similar. EDNN thus measures another aspect of semantic or form neighbourhood.

NNC: The Nearest Neighbour Correlation is an alternative way of measuring the distance of a target and its closest neighbour. Instead of the Euclidean distance, one uses the highest correlation value of the target’s predicted vector with any of the other vectors in  $\mathbf{S}$  or  $\mathbf{C}$ . The highest value identifies the nearest neighbour and estimates how close the target is to its nearest neighbour. Like EDNN, NNC can be interpreted as a measure of semantic similarity between a target and its nearest neighbour.

Not surprisingly, the semantic measures strongly correlate with each other, with the strongest correlation for L1NORM\_S and L2NORM\_S ( $\rho = 1$ ). In order to be able to better interpret these measures we implemented a principal component analysis. The first two principal components explain 86.4% of the variance. The factor loadings of the first two principal components are given in Table 3.

Table 3: Factor loadings for the first two principal components (PC1 and PC2) derived from the semantic matrix. The suffix ‘\_s’ indicates that the measures come from the semantic matrix.

	PC1_s	PC2_s
L2NORM_S	-0.24	-0.94
NEIGHDEN_S	-0.52	0.02
EDNN_S	0.44	-0.18
NNC_S	-0.52	0.01
ALC_S	-0.47	0.30

PC1\_s has rather high negative loadings on NEIGHDEN\_S, NNC\_S and ALC\_S, and a positive loading of roughly the same magnitude on EDNN\_S. This constellation can be interpreted in such a way that PC1\_s is a measure of semantic distance: compound tokens with a higher PC1\_s are characterised by a greater semantic distance to other compound tokens. PC2\_s clearly and primarily captures the length of the semantic vector, with a negative correlation between PC1\_s and L2NORM\_S (of the two length measures, we included only L2NORM\_S in the principal component analysis).

The form measures also strongly correlate with each other, and we implemented again a principal component analysis. The first two principal components explain 86.7% of the variance. The factor loadings of

the first two principal components are given in Table 4.

Table 4: Factor loadings for the first two principal components (PC1 and PC2) derived from the form matrix. The suffix ‘\_c’ indicates that the measures come from the form matrix.

	PC1_f	PC2_f
L2NORM_C	0.07	-0.73
NEIGHDEN_C	-0.59	-0.20
EDNN_C	0.32	-0.63
NNC_C	-0.57	-0.17
ALC_C	-0.47	-0.09

PC1\_c has the highest (negative) loadings on NEIGHDEN\_C, NNC\_C and ALC\_C, and a smaller positive loading on EDNN\_C. Like before, this constellation means that PC1\_c is a measure of distance between a target and other words: compound tokens with a higher PC1\_c have a pitch contour that is more different from that of other compound tokens than a token with a lower value of PC1\_c. PC2\_c has strong negative loadings on L2NORM\_C and EDNN\_C. A high value of L2NORM\_C is an indication of pronounced excursions and more oscillations in the pitch contours. As L2NORM\_C negatively correlates with PC1\_c, this means that a high PC1\_f indicates rather flat curves. Furthermore, PC1\_c also taps into EDNN\_C. Recall that EDNN\_C is the Euclidean distance of a target contour to the contours of its nearest neighbour, a high PC1\_c also means that the target’s pitch contour is rather similar to that of its nearest neighbour.

In order to see whether the LDL measures are predictive of the stress pattern we fitted a generalised linear model with the four principal components as predictors. All four principal components are significant, as shown in Table 5. To assess the power of the individual predictors we also fitted four additional models, each with only one of the four principal components, and computed the respective AICs. More powerful predictors are indicated by a smaller AIC. The first principal component of the semantic measures is the best predictor, which again indicates that semantics has a say in determining compound stress.

The partial effects are plotted in Figure 12. The two semantic predictors have negative slopes which means that being semantically more unique goes together with a lower probability of being right-stressed. This

Table 5: Logistic regression model predicting cluster membership from principal components of the LDL measures. The column labelled ‘AIC’ gives the AIC value of a regression model with this principal component as the only predictor.

	estimate	significance	standard error	AIC
(Intercept)	-0.12	n.s.	0.07	
PC1_s	-0.30	***	0.04	1287.2
PC2_s	-0.18	*	0.07	1337.1
PC1_c	0.18	***	0.05	1336.9
PC2_c	0.19	***	0.05	1329.5

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$ ; n.s. = not significant

aligns with the well-established idea that lexicalized compounds, which are often semantically idiosyncratic (e.g. *butterfly*), tend to have left stress. The form predictors have positive slopes. For PC1\_c, this means that a token with a more unique pitch contour among its neighbours is more likely to end up in cluster 2. The effect of PC2\_c shows that tokens with less extreme pitch excursions and (at the same time) a higher similarity of their pitch curve to that of its nearest neighbour’s are more prone to rightward stress.

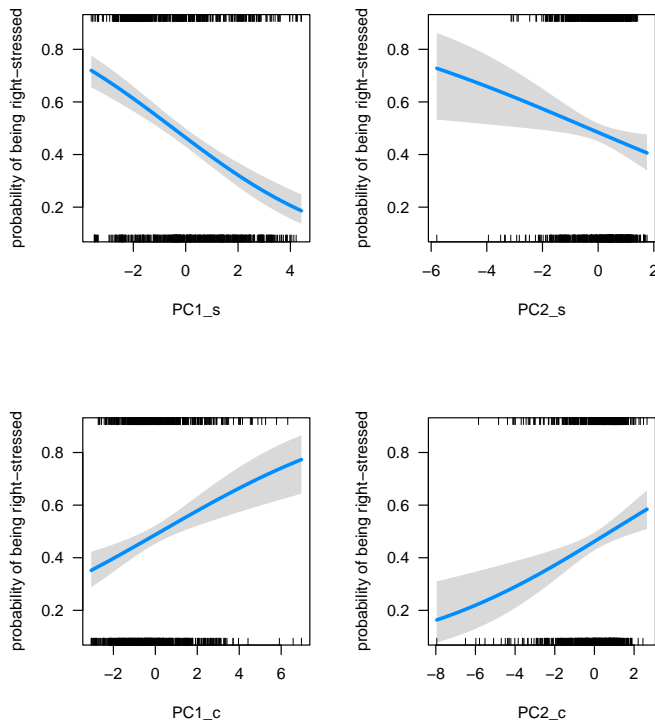


Figure 12: Partial effects of regression model predicting the stress pattern on the basis of LDL measures.

To summarise the exploration of measures extracted from the LDL model, we can say that the model is able to capture certain aspects of lexical structure that are relevant for understanding the relationship between compound semantics, pitch contours and stress patterns.

## 6 Discussion and conclusion

In this paper we investigated the prominence patterns of compounds using pitch as extracted from a speech corpus. Pitch contours were modelled using GAMs. The analysis of the pitch contours demonstrated that it is possible to group them in two sets, each of which can be interpreted as representing one of the two stress patterns advocated in large parts of the literature. The present paper is among the first studies

to provide empirical evidence for the two compound stress categories based on pitch contours (see Kösling et al. 2013; Schmitz et al. 2026 for the two others). The results are also in line with previous observations concerning within-type variability. In their experimentally elicited tokens using British English, Bell & Plag (2012) found that 37 percent of the types were variably left- or right-stressed (as rated by expert listeners). In our data, 27 percent of the types have tokens ending up in two clusters, demonstrating that within-type variability can be traced using the signal itself.

We then turned to the question of how semantics may determine the pitch contour of a compound in the context of its utterance. It was demonstrated that BERT-derived contextual embeddings encapsulate information on the semantic relation between the two constituents as well as information on the semantic category of the constituents or the compound which correlate with compound prominence as encapsulated in the pitch contour.

This is a very important finding for two reasons. First, it supports the idea that semantic aspects of a compound indeed correlate with aspects of its prominence pattern. Second, methodologically, it is now possible to circumvent questionable semantic codings of ill-defined and theoretically unsatisfactory categories or relations, and use experience-based semantic vectors instead.

Abandoning these semantic categories and semantic relations begs the question of how one can model the mapping of semantics and prominence instead. To answer that question we have tested the possibility of a direct mapping between meaning and form, using linear discriminative networks that related contextual embeddings and pitch contours to each other using multiple linear regression. The model performed well above chance when asked to find the correct type based on either the semantics (modelling comprehension) or pitch contour (modelling production) of the token in question. The discrepancy between comprehension accuracy and production accuracy can be explained by the different degree of within-type and across-type variabilities of semantic vectors vs pitch vectors. Types are more coherent semantically than formally, and types are semantically also more different from one another than formally. This makes the comprehension model more successful (in finding the right semantics) than the production model (in finding the right pitch contour).

Our findings also have implications at a higher theoretical level. In the introduction we raised the question of how speakers make use of, and learn how to use, the many different factors that are influential in determining a given compound’s prominence pattern. For linguistic theory the analogous question is what kind of theory or model can cope with the complexities of the correlations between the many different determining factors (represented by measures such as family sizes, informativity, semantic specificity, and by properties such as semantic relations and semantic categories) and prominence patterns.

Our results suggest that it is possible that variable compound prominence, i.e. variable across types and

within types, emerges from the direct mapping of form and meaning, based on the speakers' experience, which is represented as semantic vectors and form vectors in a system of discriminative learning.

That such a system is feasible also for other phenomena involving pitch contours has recently been shown in studies of Mandarin tone. Chuang et al. (2024) investigated the realization of tone in Mandarin two-character words using GAM-derived pitch contours. These authors also implemented a linear discriminative network, which was able to predict a word type based on token-specific pitch contours with 50 % accuracy on held-out data, and to predict the shape of the pitch contours based on contextualised semantic vectors with 30 % accuracy. Similar results were obtained in an analogous study of the pitch contours of monosyllabic words in Taiwan Mandarin (Jin et al., 2025). There are two things remarkable about these results.

First, they resemble the results of the present study in that the comprehension model is more successful than the production model, indicating greater variability of pitch contours as against that of contextualised embeddings across languages.

Second, the mutual predictability of meaning and pitch is not an isolated phenomenon discovered here for English compounds, but points towards a more general, direct link between meaning and form in language. Such a link is predicted by the Discriminative Lexicon Model (Baayen et al., 2019; Heitmeier et al., 2025), and studies in this framework have demonstrated, among other things, that discriminative networks using semantic vectors and form vectors other than pitch contours are able to explain intriguing durational properties of simplex and complex words (e.g. Tomaschek et al. 2021; Schmitz et al. 2021; Stein & Plag 2021; Gahl & Baayen 2024).

In the case of compounds the close link between meaning and form means that speakers, based on their experience with the compound in question or other compounds, produce pitch contours that are appropriate for the meaning they want to convey with that word, in its particular context. In comprehension, we have seen that token-specific pitch contours can indeed help listeners to find a token's meaning.

Given that meanings change slightly over different contexts, it is not surprising that one also finds clear contextual effects on compound pitch contours, as demonstrated in Schmitz et al. (2026). The direct link between experience-based semantics and compound stress patterns also explains two otherwise mysterious effects. The first being that ambiguous compounds like (*toy factory* 'a factory that makes toys', or 'a model of a factory serving as a toy') allegedly display different stress patterns (Bell & Plag, 2013). This is expected in a framework in which there is a direct link between meaning and form.

This is related to the second puzzle, namely that certain kinds of semantic categories and semantic relations go together with a tendency of the compound towards right-stress, while other semantic categories and relations tend towards left stress. Such opposite tendencies are utterly random and unclear in their provenance in a world where the pairing of meaning and form is arbitrary. In the Discriminative Lexicon

Model, the pairing of certain categories and relations with certain stress patterns would emerge naturally in a gradient fashion from experience, based on the linear mappings of particular constellations of vector dimensions on both sides, semantic and formal.

## **Data availability statement**

The data and scripts of this study are openly available in the OSF repository at [https://osf.io/gdyjr/overview?view\\_only=ac6afcc674ab4b6d8e24daaabee4d67f](https://osf.io/gdyjr/overview?view_only=ac6afcc674ab4b6d8e24daaabee4d67f).

## **Ethics statement**

Ethical approval was not required.

## **Conflict of interest statement**

The authors declare no competing interests.

## Appendix A

First, the individual pitch range of the speakers was determined. For this, the pitch information across all utterances for each of the seven speakers was extracted using Praat’s `Sound: To Pitch (raw autocorrelation)...` function (Boersma & Weenink, 2019). For female speakers, the pitch floor was set at 100 Hz and the pitch ceiling was set at 500 Hz; for male speakers, the pitch floor was set at 75 Hz and the pitch ceiling was set at 300 Hz. Then, a Praat script was used to detect voiced and unvoiced sections across all utterances (Al-Tamimi & Khattab, 2015; Al-Tamimi, 2018).<sup>10</sup> Where there was pitch detected in stretches that were found to be unvoiced, the pitch information was disregarded. This is theoretically grounded, as voiceless sounds do not have a fundamental frequency, and it is pragmatically useful, as it removed octave jumps, which mostly occurred during /s/ sounds. From the remaining pitch data, for each speaker, extreme values, i.e., values removed  $\pm 2$  standard deviations or more from the mean, were excluded on the grounds that they might represent octave jumps or other measurement errors (4.3 % of all data points excluded). The minimum and maximum pitch values of the remaining data of each speaker were then identified for use in the next step.

Second, the pitch information for the tokens was extracted using the `To Pitch (filtered ac)...` function of Praat. This was done for each speaker individually, specifying the minimum and maximum pitch values found in the previous step as pitch floor and pitch ceiling. In implementing this step, pitch floors were rounded down and pitch ceilings rounded up to the next integer (see Table 6). The `Very accurate` option of the function was turned on, resulting in more accurate pitch estimation, especially for low-pitched signals, as it applies a Gaussian window with a physical length of  $\frac{6}{pitchfloor}$  instead of a Hanning window with a physical length of  $\frac{3}{pitchfloor}$ . Pitch values were extracted at regular intervals automatically determined by Praat as  $\frac{0.75}{pitchfloor}$ , which in our data resulted in one measurement approximately every 7-10 ms depending on the speaker-specific floor value. Note that the length of the analysis window and the spacing of successive pitch measurements are independent: the window defines how much signal is used to compute each estimate, while the interval defines how frequently estimates are taken along the time axis.

Third, to get rid of octave jumps and pitch values for unvoiced sounds, the voicing information obtained with the Praat script by Al-Tamimi and Khattab (2015; 2018) was used again. The script first applied a low-pass filter at 500 Hz in order to suppress higher formants and frication noise that can interfere with reliable voicing detection. Pitch estimation was then carried out by cross-correlation in two passes: an initial run provided lower and upper quartiles of the distribution, which were used to derive a speaker-specific pitch range for the second run. From this second pass, a `PointProcess` object was created, and the mean glottal

---

<sup>10</sup>The script is available at <https://github.com/JalalAl-Tamimi/Praat-Voicing-detection>.

Table 6: Speaker-specific pitch analysis settings. The sampling interval is calculated as  $\frac{0.75}{\text{pitch floor}}$ . The speaker IDs are the ones used in the original corpus.

Speaker	Pitch floor (Hz)	Pitch ceiling (Hz)	Sampling interval (ms)
F1A	112.85	258.78	6.696
F2B	112.52	259.05	6.696
F3A	134.25	288.00	5.597
M1B	80.85	170.71	9.375
M2B	84.96	217.66	8.929
M3B	93.92	206.89	8.065
M4B	74.92	159.77	10.135

cycle length was calculated for each speaker. Praat’s voiced/unvoiced detection was subsequently applied with this mean period as input, while keeping the default minimum interval of 20 ms for continuous voiced or unvoiced stretches. This procedure differs from Praat’s regular `To Pitch...` functions in that it explicitly classifies stretches as voiced or unvoiced, rather than returning a pitch estimate for every frame possible. All pitch values that were identified as belonging to voiceless sounds were replaced with ‘NA’. That is, if an unvoiced segment within a token was previously assigned a pitch value, this pitch value was replaced with NA. Additionally, if a compound started with or ended in voiceless sounds, NAs were added at the beginning or end of its pitch measures, respectively. For example, *state officials* starts with two unvoiced sounds, i.e., /st/, which have no pitch value but are nonetheless part of the compound. Therefore, NAs were added from the onset of the word until the first pitch measure of the first vowel. The number of NAs to be added was determined by the timing between pitch measures, which in turn is dependent on the pitch floor specified for the individual speaker:  $\frac{0.75}{\text{pitch floor}}$ . The time between the onset of the compound and the first pitch measure was divided by the time in-between pitch measures. The result was the number of NAs to be added.

## References

- Al-Tamimi, J. 2018. JalalAl-Tamimi/Praat-Voicing-detection: This is the first release of version 3 of my Praat script for voicing detection. doi: 10.5281/zenodo.1183876.
- Al-Tamimi, J. & G. Khattab. 2015. Acoustic cue weighting in the singleton vs geminate contrast in Lebanese Arabic: The case of fricative consonants. *The Journal of the Acoustical Society of America* 138. 344–360. doi: 10.1121/1.4922514. URL <https://pubmed.ncbi.nlm.nih.gov/26233034/>.
- Alipoormolabashi, P. & S. Schulte im Walde. 2020. Variants of Vector Space Reductions for Predicting the Compositionality of English Noun Compounds. In *Proc. Fourth Widening Natural Language Processing Workshop*, 51–54.
- Arndt-Lappe, S. 2011. Towards an exemplar-based model of stress in English noun–noun compounds. *Journal of Linguistics* 47(3). 549–585.
- Arvaniti, A., D. R. Ladd & I. Mennen. 1998. Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of phonetics* 26(1). 3–25.
- Baayen, R. H., Y.-Y. Chuang & J. P. Blevins. 2018a. Inflectional morphology with linear mappings. *The Mental Lexicon* 13(2). 230–268. doi: <http://doi.org/10.1075/ml.18010.baa>. URL <https://www.jbe-platform.com/content/journals/10.1075/ml.18010.baa>.
- Baayen, R. H., Y.-Y. Chuang, E. Shafaei-Bajestan & J. P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in Linear Discriminative Learning. *Complexity* 2019(1). 1–39. doi: <http://doi.org/10.1155/2019/4895891>.
- Baayen, R. H., R. Piepenbrock & L. Gulikers. 1996. Celex2. *Linguistic Data Consortium, Philadelphia*.
- Baayen, R. H., M. Fasiolo, S. Wood & Y.-Y. Chuang. 2022. A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon* 17(2). 178–212.
- Baayen, R. H., J. van Rij, C. De Cat & S. Wood. 2018b. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In D. Speelman, K. Heylen & D. Geeraerts (eds.), *Mixed-effects regression models in linguistics*, 49–69. Springer.
- Bauer, L., R. Lieber & I. Plag. 2013. *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

- Bell, M. 2015a. Basic relations and stereotype relations in the semantics of compound nouns. *Journal of Cognitive Science* 16(3). 225–260.
- Bell, M. J. 2015b. Inter-speaker variation in compound prominence. *Lingue e Linguaggio* 14(1). 61–78.
- Bell, M. J. & I. Plag. 2012. Informativeness is a determinant of compound stress in English. *Journal of Linguistics* 48. 485–520.
- Bell, M. J. & I. Plag. 2013. Informativity and analogy in English compound stress. *Word Structure* 6(2). 129–155.
- Benjamin, S. & D. Schmidtke. 2023. Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory & Cognition* 1–28.
- Boersma, P. & D. Weenink. 2019. Praat: Doing Phonetics by Computer.
- Caliński, T. & J. Harabasz. 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics* 3(1). 1–27. doi: 10.1080/03610927408827101.
- Chuang, Y.-Y. & R. H. Baayen. 2021. Discriminative Learning and the lexicon: NDL and LDL. In *Oxford research encyclopedia of linguistics*, Oxford: Oxford University Press.
- Chuang, Y.-Y., M. J. Bell, Y.-H. Tseng & R. H. Baayen. 2024. Word-specific tonal realizations in Mandarin. *arXiv preprint arXiv:2405.07006* .
- Collier, R. 1975. Physiological correlates of intonation patterns. *The Journal of the Acoustical Society of America* 58(1). 249–255.
- Fanselow, G. 2011. *Zur Syntax und Semantik der Nominalkomposition: ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung im Deutschen*, vol. 107. Walter de Gruyter.
- Farnetani, E., C. T. Torsello & P. Cosi. 1988. English compound versus non-compound noun phrases in discourse: an acoustic and perceptual study. *Language and Speech* 31(2). 157–180.
- Friedman, J., T. Hastie & R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1). 1–22. doi: 10.18637/jss.v033.i01.
- Fudge, E. 1984. *English word-stress*. London: George Allen & Unwin.
- Gahl, S. & R. H. Baayen. 2024. Time and thyme again: Connecting English spoken word duration to models of the mental lexicon. *Language* 100(4). 623–670.

- Genolini, C., X. Alacoque, M. Sentenac & C. Arnaud. 2015. kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software* 65(4). 1–34. doi: 10.18637/jss.v065.i04. URL <https://www.jstatsoft.org/article/view/v065i04>.
- Giegerich, H. J. 2004. Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics* 8. 1–24.
- Günther, F. & M. Marelli. 2021. CAOSS and transcendence: Modeling role-dependent constituent meanings in compounds. *Morphology* 1–24. doi: 10.1007/s11525-021-09386-6.
- Günther, F. & M. Marelli. 2022. Patterns in CAOSS: Distributed representations predict variation in relational interpretations for familiar and novel compound words. *Cognitive Psychology* 134. 101471.
- Günther, F., M. A. Petilli & M. Marelli. 2020. Semantic transparency is not invisibility: A computational model of perceptually-grounded conceptual combination in word processing. *Journal of Memory and Language* 112. 104104.
- Gussenhoven, C., B. H. Repp, A. Rietveld, H. H. Rump & J. Terken. 1997. The perceptual prominence of fundamental frequency peaks. *The Journal of the Acoustical Society of America* 102(5). 3009–3022.
- Gussenhoven, C. & A. C. M. Rietveld. 1988. Fundamental frequency declination in Dutch: testing three hypotheses. *Journal of phonetics* 16(3). 355–369.
- Heitmeier, M. 2022. JudiLingMeasures.jl. <https://github.com/MariaHei/JudiLingMeasures.jl>.
- Heitmeier, M., Y.-Y. Chuang & R. H. Baayen. 2025. *The discriminative lexicon: Theory and implementation in the Julia package JudiLing*. Cambridge University Press.
- Ingram, J. & T. Nguyen. 2007. Prosodic typology and compound–phrasal contrasts. In *Proceedings of the 15th international congress of phonetic sciences*, 479–482. Barcelona.
- Jin, X., M. Ernestus & R. H. Baayen. 2025. The New Kid on the Block: The Role of Word Meaning in the Realization of Tone in Conversational Taiwan Mandarin Speech. *Available at SSRN 5304595* .
- Kamin, L. J. 1969. Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (eds.), *Punishment and aversive behavior*, 276–296. New York: Appleton-Century-Crofts.
- Kösling, K., G. Kunter, R. H. Baayen & I. Plag. 2013. Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech* 56(4). 529–554.

- Kunter, G. 2010. Perception of prominence patterns in English nominal compounds. *Proceedings of Speech Prosody 2010* 102007. 1–4.
- Kunter, G. 2011. *Compound stress in English: The phonetics and phonology of prosodic prominence*. Berlin & Boston: De Gruyter Mouton.
- Kunter, G. & I. Plag. 2007. What is compound stress? In J. Trouvain & W. J. Barry (eds.), *Proceedings of the 16th international congress of phonetic sciences*, 1005–1008. Saarbrücken, Germany.
- Ladd, D. R., J. Verhoeven & K. Jacobst. 1994. Influence of adjacent pitch accents on each other’s perceived prominence: Two contradictory effects. *Journal of phonetics* 22(1). 87–99.
- Levi, J. N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.
- Lieberman, A. M., K. S. Harris, H. S. Hoffman & B. C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology* 54(5). 358.
- Lieberman, M. Y. & R. Sproat. 1992. The stress and structure of modified noun phrases in English. In I. A. Sag & A. Szabolcsi (eds.), *Lexical Matters*, 131–181. Stanford: CSLI publications.
- Lieber, R. 2021. *Introducing morphology*. Cambridge University Press.
- van der Maaten, L. & G. Hinton. 2008. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* 9(86). 2579–2605.
- Marelli, M., C. L. Gagné & T. L. Spalding. 2017. Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition* 166. 207–224.
- Mikolov, T., K. Chen, G. Corrado & J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* doi: 10.48550/arxiv.1301.3781.
- Mitchell, J. & M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8). 1388–1429.
- Nieder, J., Y.-Y. Chuang, R. van de Vijver & R. H. Baayen. 2022. A Discriminative Lexicon approach to word comprehension, production and processing: Maltese plurals. Accepted for publication. *Language* URL <https://psyarxiv.com/rkath/>.

- Ó Séaghdha, D. 2008. Learning compound noun semantics. Tech. rep. University of Cambridge, Computer Laboratory Cambridge.
- OED. 2022. The Oxford English Dictionary online: [www.oed.com](http://www.oed.com).
- Ostendorf, M., P. Price & S. Shattuck-Hufnagel. 1996. *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium.
- Pearce, J. M. & M. E. Bouton. 2001. Theories of associative learning in animals. *Annual review of psychology* 52(1). 111–139.
- Petilli, M. A., F. Günther, A. Vergallito, M. Ciapparelli & M. Marelli. 2019. Data-driven computational models reveal perceptual simulation in word processing. *psyarxiv.com* .
- Plag, I. 2006. The variability of compound stress in English: Structural, semantic, and analogical factors. *English Language and Linguistics* 10(1). 143–172.
- Plag, I. 2010. Compound stress assignment by analogy: the constituent family bias. *Zeitschrift für Sprachwissenschaft* 29(2). 243–282.
- Plag, I. 2018. *Word-formation in English, 2nd edition*. Cambridge: Cambridge University Press.
- Plag, I., M. Heitmeier & F. Domahs. 2025. Morpho-phonology is not independent of semantics: The case of German nominal number marking. *The Mental Lexicon* .
- Plag, I. & G. Kunter. 2010. Constituent family size and compound stress assignment in English. *Linguistische Berichte Sonderheft 17*. 349–382.
- Plag, I., G. Kunter & S. Lappe. 2007. Testing hypotheses about compound stress assignment in English: A corpus-based investigation. *Corpus Linguistics and Linguistic Theory* 3(2). 199–233.
- Plag, I., G. Kunter, S. Lappe & M. Braun. 2008. The role of semantics, argument structure, and lexicalization in compound stress assignment in English. *Language* 84(4). 760–794.
- Plag, I., G. Kunter & M. Schramm. 2011. Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics* 39(3). 362–374.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL <https://www.R-project.org/>.

- Ramscar, M., D. Yarlett, M. Dye, K. Denny & K. Thorpe. 2010. The Effects of feature-label-order and their implications for symbolic learning. *Cognitive Science* 34(6). 909–957. doi: 10.1111/j.1551-6709.2009.01092.x.
- Repp, B. H. 1984. Categorical perception: Issues, methods, findings. In *Speech and language*, vol. 10, 243–335. Elsevier.
- Rescorla, R. A. 1988a. Behavioral studies of Pavlovian conditioning. *Annual Review of Neuroscience* 11(1). 329–352.
- Rescorla, R. A. 1988b. Pavlovian conditioning. It’s not what you think it is. *American Psychologist* 43(3). 151–160. doi: 10.1037/0003-066X.43.3.151.
- Rescorla, R. & A. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (eds.), *Classical conditioning II: Current research and theory*, 64–99. New York: Appleton-Century-Crofts.
- Rietveld, A. C. & C. Gussenhovent. 1985. On the relation between pitch excursion size and prominence. *Journal of phonetics* 13(3). 299–308.
- Schäfer, M. & M. J. Bell. 2020. Constituent polysemy and interpretational diversity in attested English novel compounds. *The Mental Lexicon* 15(1). 42–61.
- Schakel, A. M. & B. J. Wilson. 2015. Measuring word significance using distributed representations of words. *arXiv preprint arXiv:1508.02297* .
- Schmitz, D., M. Bell & I. Plag. 2026. Compound prosody in context: The influence of speaker, compound type and context on the pitch contours of noun-noun compounds. In J. Nieder & I. Plag (eds.), *Morphological variation*, De Gruyter.
- Schmitz, D., I. Plag, D. Baer-Henney & S. Stein. 2021. Durational differences of word-final /s/ emerge from the lexicon: Modeling morpho-phonetic effects in pseudowords with Linear Discriminative Learning. *Frontiers in Psychology* 12. doi: 10.3389/fpsyg.2021.680889. 680889.
- Schmitz, D., V. Schneider & J. Esser. 2023. No Genericity in Sight: An Exploration of the Semantics of Masculine Generics in German. *Glossa Psycholinguistics* 2(1). doi: 10.5070/G6011192.
- Schulte im Walde, S., A. HäTTY & S. Bott. 2016. The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective. In *Proc. 5th Joint Conf. on Lexical and Computational Semantics*, 148–158. Berlin, Germany: ACL.

- Simon, N., J. Friedman, T. Hastie & R. Tibshirani. 2011. Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 39(5). 1–13. doi: 10.18637/jss.v039.i05.
- Stein, S. D. & I. Plag. 2021. Morpho-phonetic effects in speech production: Modeling the acoustic duration of English derived words with Linear Discriminative Learning. *Frontiers in Psychology* 12. doi: 10.3389/fpsyg.2021.678712. 678712.
- Tay, J. K., B. Narasimhan & T. Hastie. 2023. Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software* 106(1). 1–31. doi: 10.18637/jss.v106.i01.
- Terken, J. 1997. Variation of accent prominence within the phrase: Models and spontaneous speech data. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*, 95–111. Springer.
- Terken, J. & D. Hermes. 2000. The perception of prosodic prominence. In *Prosody: Theory and experiment: Studies presented to Gösta Bruce*, 89–127. Springer.
- Tomaschek, F., I. Plag, M. Ernestus & R. H. Baayen. 2021. Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics* 57(1). 123–161.
- Van de Vijver, R. & E. Uwambayinema. 2022. A word-based account of comprehension and production of Kinyarwanda nouns in the Discriminative Lexicon. *Linguistics Vanguard* 8(1). 197–207.
- Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest & A. M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- Wood, S. N. 2017. *Generalized additive models: An introduction with R*. CRC Press. doi: 10.1201/9781315370279. URL <https://www.taylorfrancis.com/books/mono/10.1201/9781315370279/generalized-additive-models-simon-wood><https://www.taylorfrancis.com/books/9781498728348>.
- Zwicky, A. M. 1986. Forestress and afterstress. In *Working Papers in Linguistics*, vol. 32, 46–72. Columbus: Ohio State University.