Gradient Selectional Restrictions in HPSG

Adam Przepiórkowski (ICS Polish Academy of Sciences and University of Warsaw)
Berke Şenşekerci (University of Warsaw)

Introduction Syntactic selectional restrictions are often – usually implicitly – assumed to be strict. For example, *wait* selects for a PP (headed by *for* or *on*), e.g., *Homer waited for Marge*, while the near-synonymous *await* selects for an NP, e.g., *Homer awaited Marge*, and violating such selectional restrictions is assumed to lead to ungrammaticality, e.g., **Homer waited Marge* or **Homer awaited for Marge*.

In this paper, we first show experimentally that syntactic selectional restrictions may be gradient – in the sense that their violation leads to relatively mild unacceptability, rather than to strong unacceptability that may reflect ungrammaticality. As we do not find an explanation for this mild unacceptability in terms of processing, we conjecture that it should be modelled directly in the grammar. Accordingly, the second aim of this paper is to show how such gradient grammaticality of selectional restrictions may in principle be modelled in HPSG. In this part, we build on a recent proposal to include gradience in the HPSG model theory.

Experiment For the experiment, we selected 8 predicates that combine with PPs, namely: *account (for)*, *annoyed (by)*, *ashamed (of)*, *depend (on)*, *familiar (with)*, *speak (about)*, *suffer (from) talk (about)*. Each of these predicates is semantically compatible with a proposition syntactically expressed by a PP, e.g., *account for the fact that…* or *familiar with the fact that…*. The aim of this experiment was to test to what extent this proposition may be syntactically realised as a CP, e.g., *account that…* or *familiar that…*. Hence, the experiment reported here had a 8×2 design: for each of the 8 predicates, 2 types of dependents were considered: PPs and CPs.

For example, one of the items ("tokensets"; Cowart 1997) for *depend* was the following pair:[1]

(1) a. Justin Bieber can depend [$_{\text{PP}}$ on the fact [$_{\text{CP}}$ that his fans still love his early songs]].
b. Justin Bieber can depend [$_{\text{CP}}$ that his fans still love his early songs].

In (1a), *depend* combines with the PP *on the fact that…*, while in (1b) it combines with the embedded CP directly. For each predicate, there were eight such minimal pairs, so there were $8 \times 8 = 64$ pairs like (1), i.e., $2 \times 64 = 128$ stimuli altogether. The experiment followed the usual Latin square design, so that – for each predicate – each participant judged PP stimuli like (1a) from four tokensets and CP stimuli like (1b) from the other four tokensets.

We measured acceptability using the Thermometer Method (TM; Featherston 2008), argued in Featherston 2009 to combine the strengths of Likert Scale and Magnitude Estimation (ME; Bard *et al.* 1996), while avoiding their weaknesses. The idea of TM is that respondents are presented with two sentences, visible throughout the experiment: one which is relatively unacceptable (but interpretable) and one that is relatively acceptable (but perhaps not completely natural). To this end, we used the following two sentences from Featherston 2021: 53:

(2) The father fetches the wholemeal bread them.
(3) The father fetches the children the wholemeal bread.

Sentence (2) is assigned value 20 and (3) is assigned 30, and respondents are asked to judge other sentences by comparing them to these two sentences. The expectation is that the vast majority of judgements will fall within the range of 15–35 (Featherston 2009: 64), so the scale in the experiment was limited to 10–40. (However, around 25% of all responses were outside of the 15–35 range, so we use the whole 10–40 range in the model below.)

In order to better determine the absolute acceptability, we included among the fillers the standard items from Gerbrich *et al.* 2019, i.e., sentences grouped into 5 classes, A–E, shown to reliably produce the whole range of different acceptability ratings.[2] Each class consists of 3 sentences; those in classes A and B are grammatical (e.g., *The winter is very harsh in the North.* and *Before every lesson the teacher must prepare their materials.*, respectively), C are borderline (e.g., *Anna loves, but Linda hates, eating popcorn at the cinema.* and *Most people like very much a cup of tea in the morning.*), and D and E are ungrammatical (e.g., *The old fisherman took her pipe out of mouth and began story.* and *Student must read much book for they become clever.*, respectively).

89 native speakers of English completed the experiment, but 19 were excluded from analysis due to failing one or more of the strict exclusion criteria (e.g., participants had to answer at least 9 out of 12 comprehension questions correctly, they could not be current or former students of linguistics, etc.). The results of the experiment are summarized in Figure 1, which presents mean z-scores of the stimuli – grouped by the predicate and the PP vs. CP variant – and standard items.[3]

[1] Brackets and categories are added for perspicuity here; they were absent in actual experimental items.

[2] For example, on the 1–5 Likert Scale, examples in class A produce average ratings close to 5, B – around 4, C – a little above 3, D – a little above 2, and E – between 1 and 1.5 (Gerbrich *et al.* 2019: 314–315; cf. Molimpakis 2019: 128–130, Featherston 2020: 169–171, and Brown *et al.* 2021: 16–18).

[3] All statistical analyses were performed using R (R Core Team 2024). As is common, we used z-scores to eliminate scale effects.

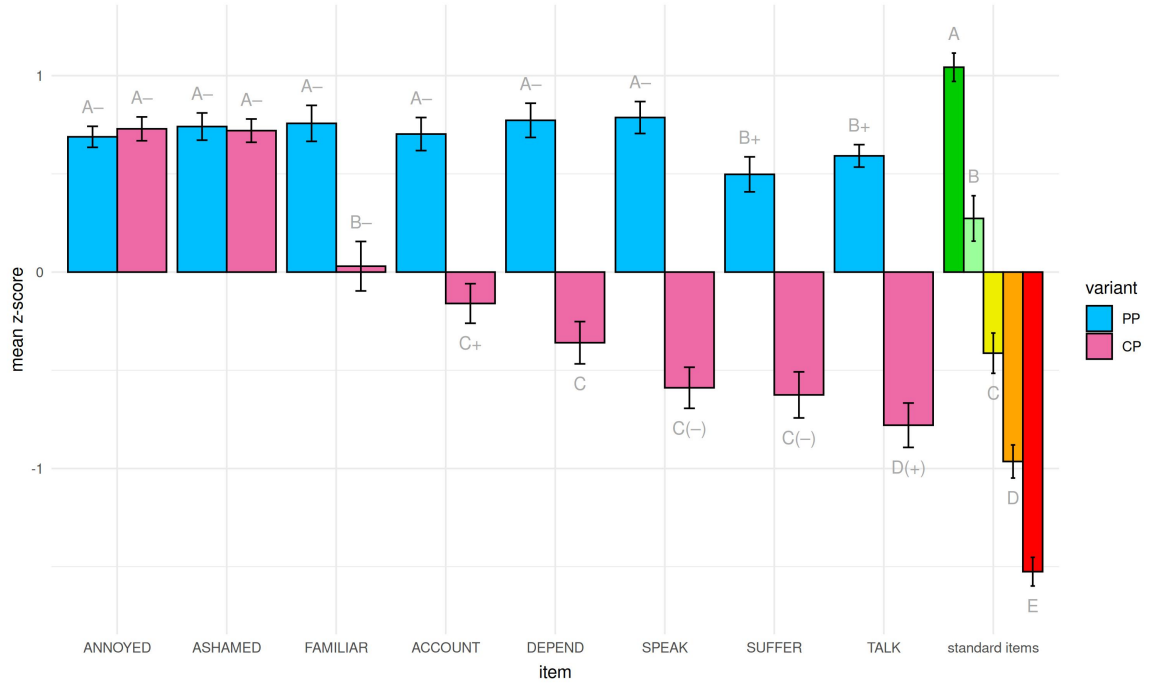

Figure 1: Average z-scores of stimuli and standard items (with 95% confidence intervals)

While the acceptability of PP variants was always high (between A− and B+, in terms of standard items), the acceptability of CP variants varied greatly: from fully acceptable *annoyed* and *ashamed* (A−), through marginally acceptable *familiar* (B−) and *account* (C+) and a little less acceptable *depend* (C), to the least acceptable *speak* (C(−)), *suffer* (C(−)), and *talk* (D(+)).[4] According to linear mixed effects models and pairwise comparisons (with Tukey's HSD adjustment for multiple comparisons),[5] only for *annoyed* and *ashamed* was there no statistically significant difference in acceptability between PP and CP sentences; for the other six predicates, all differences were statistically significant ($p < .001$). On the other hand, all means of CP conditions – with the sole exception of *talk* – were statistically significantly higher not only than that of standard items E, but also than that of standard items D, so even *depend* or *speak* – never mind *familiar* or *account* – do not seem to be ungrammatical with CP complements.

In summary, predicates similarly happy to combine with propositional content when it is expressed by a PP differ significantly in their readiness to combine with a declarative CP. As CPs are prototypical syntactic categories of propositions, this difference does not seem to be of semantic nature. Moreover, we controlled for length and complexity, so it is unlikely that these differences reflect some processing differences. In short, we do not know how to explain this observation other than by admitting that syntactic selectional restrictions may be gradient within grammar proper.

Gradient Grammaticality We propose modelling gradience of selectional restrictions within the gradient HPSG framework outlined in Şenşekerci 2024 (with minor modifications below). On this approach, HPSG grammatical constraints (i.e., RSRL formulae; Richter 2004) are assigned weights. Formally, an HPSG theory θ on this view is not just a set of closed formulae δ_i (Richter 2024: 173), but a set of $\langle \delta_i, w_i \rangle$ pairs such that δ_i is a formula and $w_i \in [0, 1]$ is its weight – the cost of a single violation of this formula.

If all model objects satisfy all constraints, as usually assumed in HPSG, then the weight of the model is zero. However, if some objects violate some constraints, then the weight of the model is the sum of the weights of the violated constraints (Keller 2000; cf. fn. 7 below). (Violating the same constraint multiple times also has a roughly additive effect.) Hence, each model $\mathfrak{I}$ comes with a burden B($\mathfrak{I}$) calculated as in (4):[6]

[4] We used the following convention for assigning acceptability marks such as "B+" or "C(−)" to means of particular items. The initial letter (e.g., "B") is that of the standard items with means closest to the means of the current items. If half or more of the confidence interval of the current items is within that of the closest standard items, then the letter remains unadorned (here, "B"). Otherwise, if the two confidence intervals overlap, then the letter is adorned with "(+)" or "(−)", depending on the direction of this overlap (resulting, e.g., in "C(−)"). Otherwise, i.e., if there is no overlap of confidence intervals, the letter is adorned with "+" or "−" (e.g., "B+").

[5] The standard *lme4* (Bates *et al.* 2015) and *emmeans* (Lenth 2024) R packages were used here.

[6] An important assumption behind this approach is that models correspond to single utterances, as often assumed in model-theoretic syntax and as proposed for HPSG in Przepiórkowski 2021: §2; this is unlike in standard RSRL, which implements the view of King 1999 that models should be exhaustive, i.e., that each should contain configurations of *all* actual and potential utterances in a language.

(4) $\mathrm{B(I)} = \sum_{\langle \delta_i, w_i \rangle \in \theta} |\mathsf{U} \backslash \mathsf{D_I}(\delta_i)| \cdot w_i$,

where U is the set of all model objects and $\mathsf{D_I}(\delta_i)$ is the set of model objects satisfying the constraint δ_i, so $|\mathsf{U} \backslash \mathsf{D_I}(\delta_i)|$ is the number of objects *not* satisfying δ_i. Hence, the formula in (4) has the effect of adding to the model's burden the weight w_i each time δ_i is violated, for all constraints δ_i.[7] Of course, models with smaller burdens – without constraint violations or with very few violations of very "light" constraints – represent more grammatical structures than models with higher burden, in which more and/or "heavier" constraints are violated.

Constraining Selectional Restrictions Given that, in standard HPSG, multiple selectional restrictions are expressed within a single lexical entry, and that such entries are sometimes assumed to be just small parts of a single constraint defining the lexicon (the Word Principle – see, e.g., Richter 2024: 101), it is far from clear how to attach a weight to a single selectional restriction of a single lexical item. We consider two options below.

Option 1 Assume that some selectional restrictions are hard and others are soft (cf. Bresnan *et al.* 2001). For example, a hard constraint on the complement of *depend* could be that it must be a PP[*on*] (or PP[*upon*], but we do not model register here) or a CP[*that*] (we ignore interrogative CPs here), but it cannot be an NP:

(5) *Justin Bieber can depend [$_{\mathrm{NP}}$ the fact [$_{\mathrm{CP}}$ that his fans still love his early songs]].

If so, lexical entries could be taken to express such hard constraints, i.e., they would only mention argument realisations that lead to at most soft constraint violations, as in the lexical entry for *depend* in (6).[8]

(6) lexical entry: $\begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \text{NP}[\textsc{nom}], \text{PP}[\textit{on}] \vee \text{CP}[\textit{that}] \rangle \end{bmatrix}$

Assuming that the Word Principle (or any other lexicon-defining constraint) that contains the lexical entry in (6) has the maximal weight 1, this lexical entry ensures that the second argument of *depend* cannot be an NP. However, it says nothing about the respective weights of the two allowed options PP[*on*] and CP[*that*]. For this, separate constraints with individual weights are needed, such as (7), which says that the second argument of *depend* cannot be a CP, or the penalty of, say, 0.28 will be imposed on the model. As there is no similar constraint forbidding this argument to be a PP, realising it as a PP[*on*] incurs no costs.

(7) $\left\langle \begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2} \rangle \end{bmatrix} \rightarrow \neg \boxed{2}\text{CP},\ 0.28 \right\rangle$

Option 2 Assume now that gradience permeates grammar, i.e., that there is no strict difference between soft and hard constraints; hard constraints are simply those whose weights are relatively close to 1. If so, all syntactic selectional restrictions should be stated in separate constraints, with information in lexical entries limited to the number of arguments (which itself might follow from semantics rather than being stated explicitly), see (8).

(8) lexical entry: $\begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2} \rangle \end{bmatrix}$

In order to attach weights to all possible morphosyntactic realisations of a given argument, we propose two tiers of constraints: coarse-grained category constraints and fine-grained morphosyntactic constraints. Coarse category constraints may be defined in terms of maximal subsorts of *head*, as in (9)–(11), etc. These constraints allow in principle for PP realisations of the second argument of *depend*, impose a soft constraint on CP realisations (assumed here to be headed by complementisers), and practically exclude NP realisations, etc.

(9) $\left\langle \begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2} \rangle \end{bmatrix} \rightarrow \neg \boxed{2}\big[\ldots | \textsc{head}\ \textit{preposition}\big],\ 0.00 \right\rangle$

(10) $\left\langle \begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2} \rangle \end{bmatrix} \rightarrow \neg \boxed{2}\big[\ldots | \textsc{head}\ \textit{complementiser}\big],\ 0.28 \right\rangle$

(11) $\left\langle \begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2} \rangle \end{bmatrix} \rightarrow \neg \boxed{2}\big[\ldots | \textsc{head}\ \textit{noun}\big],\ 0.93 \right\rangle$

However, not all PPs are possible, only PP[*on*]s (we keep ignoring PP[*upon*]). Similarly, let us assume that not all CPs are possible as direct complements of *depend*, only declarative CP[*that*]s. So, while more fine-grained specifications of the practically excluded NP option are not needed, PP and CP must be further specified, e.g.:

(12) $\left\langle \begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2}\big[\ldots | \textsc{head}\ \textit{preposition}\big] \rangle \end{bmatrix} \rightarrow \boxed{2}\big[\ldots | \textsc{head} | \textsc{pform}\ \textit{on}\big],\ 0.97 \right\rangle$

(13) $\left\langle \begin{bmatrix} \textsc{phon} \langle \text{depend} \rangle \\ \ldots \textsc{arg-st} \langle \boxed{1}, \boxed{2}\big[\ldots | \textsc{head}\ \textit{complementiser}\big] \rangle \end{bmatrix} \rightarrow \boxed{2}\big[\ldots | \textsc{head} | \textsc{compform}\ \textit{that}\big],\ 0.91 \right\rangle$

The combined effect of (9) and (12) is that the relevant argument may be a PP[*on*] but – at the penalty of 0.97 – not any other kind of PP, and similarly for the combined effect of (10) and (13), according to which CP[*that*] incurs the penalty of 0.28, while any other CP would add 0.91 to that penalty.

[7] Given the sometimes reported *under*additive effect of multiple violations, the formula in (4) must be more complex, but the current psycholinguistic evidence does not make it possible to make this formula more precise here. An example of a formula that would model the underadditive effect of constraint violations and keep the model burden within the $[0, 1]$ range is: $\mathrm{B(I)} = 1 - \prod_{\langle \delta_i, w_i \rangle \in \theta} (1 - w_i)^{|\mathsf{U} \backslash \mathsf{D_I}(\delta_i)|}$. For example, three weighty violations worth 0.7, 0.8, and 0.9 would result in a very heavy but *strongly under*additive burden equal $1-(1-0.7)\cdot(1-0.8)\cdot(1-0.9) = 0.994 \approx 1.0$ (the "floor effect" in acceptability judgements), while a double violation of a very light constraint with weight equal to 0.05 would result in the *almost fully* additive small burden $1-(1-0.05)^2 = 0.0975 \approx 0.1$.

[8] See Yatabe 2004 and Przepiórkowski 2021: §4 on such disjunctive selectional restrictions (and their interpretation in coordination).

Weights A central issue in the above proposal is the origin and meaning of weights. A methodologically cautious view is that weights reflect the impact of a constraint violation as expressed on a given experimental scale by participants, and nothing more. On this view, weights are essentially understood as modelling devices: they help us quantify how far an anomalous sentence is from full well-formedness in a given language. A stronger interpretation would go further and treat weights as neurocognitively real – i.e., as directly reflecting the magnitude of responses in neural systems that register constraint violations. While this possibility is appealing, it remains an open empirical question that requires closer integration of behavioural and neuroimaging evidence. For present purposes, we adopt the more conservative view and treat weights as descriptive measures derived from judgement data.

Regardless of how weights are ultimately interpreted, a uniform procedure is needed to extract them from judgement data. Şenşekerci (2024: §4.4) proposes a method in which weights are equated with fixed effects coefficients from mixed-effects models, building on the multiple linear regression approach of Keller (2000: §6.3.4). However, coefficients obtained in this way are tied to the specific scale used in a given experiment and are therefore not directly comparable across studies. Ideally, a weight extraction procedure should yield values that generalize across experiments, but this is difficult to achieve for several reasons.

First, acceptability judgement studies employ different response scales, ranging from bounded Likert scales to unbounded methods such as Magnitude Estimation (Bard *et al.* 1996, Cowart 1997). Second, even when the same scale is used, participants vary in how they use it, and this variation is itself sensitive to the overall acceptability of materials in the experiment, including fillers (Häussler and Juzek 2021: 108). Third, there are different ways of operationalizing weights. Assuming that a weight reflects the impact of a violation, estimating this impact crucially depends on the choice of baseline against which it is measured. One can take the fully grammatical control condition as baseline (e.g., PP conditions in the experiment described above), or alternatively a set of highly acceptable filler items (e.g., A-level standard items in the experiment described above), and these choices are not guaranteed to yield equivalent results.

Against this backdrop, we propose a simple pipeline that stays as close as possible to standard practice in the analysis of judgement data. We begin by fitting a mixed-effects model to the raw, untransformed ratings, with participants and items included as random effects and the relevant conditions as fixed effects.[9] A key decision at this stage concerns the choice of baseline. Here we take the fully grammatical control condition as the reference level (i.e., the PP condition). Under this choice, for example, the estimated drop in acceptability from PP to CP in *depend* items is approximately 8.44 points on the 10–40 scale.

While this estimate incorporates both participant- and item-level variation, this estimate is meaningful only relative to the scale used in the experiment. As such, it cannot be directly compared to estimates obtained under different scaling regimes. To address this, we rescale coefficient estimates post hoc using min–max normalization.[10] We define a weight w as the absolute coefficient normalized by the total range of the experimental scale, as in (14). Here β is the model coefficient associated with the violation (e.g., $\beta = -8.44$ for DEPEND_CP). Importantly, the scale bounds in the denominator can be defined either in terms of the theoretical limits of the provided scale or the observed range. In the present case, the observed range was the same as the full theoretical range 10–40 (around 6.5% of responses were 40 and around 1.5% were 10), so $S_{max} = 40$ and $S_{min} = 10$. Applying this procedure to four of the eight predicates in the experiment described above yields the weights in (15)–(18).

(14) $w_{violation} = \frac{|\beta_{violation}|}{S_{max} - S_{min}}$

(15) $w_{\text{ASHAMED\_CP}} = \frac{|-0.27|}{40-10} = 0.01$

(16) $w_{\text{FAMILIAR\_CP}} = \frac{|-5.58|}{40-10} = 0.19$

(17) $w_{\text{DEPEND\_CP}} = \frac{|-8.44|}{40-10} = 0.28$

(18) $w_{\text{TALK\_CP}} = \frac{|-10.04|}{40-10} = 0.33$

These values can be interpreted straightforwardly: for instance, a weight of 0.28 for DEPEND_CP indicates an average loss of approximately 28% of the observed scale range relative to the PP baseline. For ASHAMED_CP, the reduction is practically zero, and for TALK_CP, the reduction is 33%. In conclusion, weights derived this way provide a normalized measure of violation strength that is comparable across items and, in principle, across experiments that differ in their scaling conventions.

Earlier Work We are not aware of any earlier work formalising gradient selectional restrictions. The proposal above is different from "probabilistic" work (e.g., Probabilistic CFGs, Brew 1995, etc.) as it models acceptability, rather than corpus probabilities (see, e.g., Featherston 2005 on differences between the two). More directly related are the computational frameworks Weighted Constraint Dependency Grammar (Schröder *et al.* 2000) and (weighted) Property Grammars (Blache *et al.* 2006), which however say nothing about the origin and meaning of weights attached to constraints or about gradient selectional restrictions.

[9] For unbounded scales such as Magnitude Estimation, an appropriate transformation would be required prior to modelling.

[10] While non-linear functions can be used for scaling, they typically presuppose non-linear relationships between scale points. The min–max approach is preferred here as it remains agnostic about such perceptual compression and preserves the raw linear intervals of the judgement data.

References

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

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Blache, P., Hemforth, B., and Rauzy, S. (2006). Acceptability prediction by means of grammaticality quantification. In N. Calzolari, C. Cardie, and P. Isabelle, eds., *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 57–64. Association for Computational Linguistics.

Bresnan, J., Dingare, S., and Manning, C. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt and T. H. King, eds., *Proceedings of the LFG'01 Conference*, pp. 13–32. CSLI Publications.

Brew, C. (1995). Stochastic HPSG. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1995)*, pp. 83–89.

Brown, J. M. M., Fanselow, G., Hall, R., and Kliegl, R. (2021). Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm. *PLoS ONE*, **16**(5), 1–26.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications.

Featherston, S. (2005). The Decathlon model of empirical syntax. In S. Kepser and M. Reis, eds., *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, pp. 187–208. Mouton de Gruyter.

Featherston, S. (2008). Thermometer judgements as linguistic evidence. In C. M. Riehl and A. Rothe, eds., *Was ist linguistische Evidenz?*, pp. 69–90. Shaker Verlag.

Featherston, S. (2009). A scale for measuring well-formedness: Why syntax needs boiling and freezing points. In S. Featherston and S. Winkler, eds., *The Fruits of Empirical Linguistics (Volume 1: Process)*, pp. 47–73. Mouton de Gruyter.

Featherston, S. (2020). Can we build a grammar on the basis of judgments? In S. Schindler, A. Drożdżowicz, and K. Brøcker, eds., *Linguistic Intuitions: Evidence and Method*, pp. 165–188. Oxford University Press.

Featherston, S. (2021). Response methods in acceptability experiments. In Goodall (2021), pp. 39–61.

Gerbrich, H., Schreier, V., and Featherston, S. (2019). Standard items for English judgment studies: Syntax and semantics. In S. Featherston, R. Hörnig, S. von Wietersheim, and S. Winkler, eds., *Experiments in Focus: Information Structure and Semantic Processing*, pp. 305–327. Mouton de Gruyter.

Goodall, G., ed. (2021). *The Cambridge Handbook of Experimental Syntax*. Cambridge University Press.

Häussler, J. and Juzek, T. S. (2021). Variation in participants and stimuli in acceptability experiments. In Goodall (2021), pp. 97–117.

Keller, F. (2000). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh.

King, P. J. (1999). Towards truth in HPSG. In V. Kordoni, ed., *Tübingen Studies in Head-Driven Phrase Structure Grammar*, pp. 301–352. Universität Tübingen.

Lenth, R. V. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.0.

Molimpakis, E. (2019). *Accepting Preposition-Stranding under Sluicing Cross-linguistically; a Noisy-Channel Approach*. Ph.D. dissertation, UCL.

Przepiórkowski, A. (2021). Three improvements to the HPSG model theory. In S. Müller and N. Melnik, eds., *Proceedings of the HPSG 2021 Conference*, pp. 165–185. Frankfurt/Main University Library.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Richter, F. (2004). *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar*. Ph.D. dissertation (defended in 2000), Universität Tübingen.

Richter, F. (2024). Formal background. In S. Müller, A. Abeillé, R. D. Borsley, and J.-P. Koenig, eds., *Head-Driven Phrase Structure Grammar: The Handbook*, pp. 93–131. Language Science Press, 2nd edition.

Schröder, I., Menzel, W., Foth, K., and Schulz, M. (2000). Modeling dependency grammar with restricted constraints. *Traitement Automatique des Langues (T.A.L.)*, **41**(1), 97–126.

Şenşekerci, B. (2024). Gradient HPSG. In S. Müller and R. Chaves, eds., *Proceedings of the HPSG 2024 Conference*, pp. 173–192.

Yatabe, S. (2004). A comprehensive theory of coordination of unlikes. In S. Müller, ed., *Proceedings of the HPSG 2004 Conference*, pp. 335–355. CSLI Publications.