Evaluating Unalternative Semantics on a large corpus of naturalistic speech

Fenna Berger
Leiden University / Vrije Universiteit Amsterdam
Leiden / Amsterdam, The Netherlands
f.berger@student.vu.nl

Matthijs Westera
LUCL, Leiden University
Leiden, The Netherlands
m.westera@hum.leidenuniv.nl

Abstract

While the experimental turn of the last decade has helped connect formal semantic/pragmatic theories to empirical reality, extending this empirical reach to naturally-occurring data has lagged behind. One of the challenges is the lack of large corpora in which the relevant technical notions have been annotated. Nevertheless, some insight may be gained by relying on automatically computed approximations of these notions. We take such an approach to focus realization – the relation between focus, grammar and pitch accents, taking inspiration in particular from Unalternative Semantics.

1 Introduction

Unalternative Semantics (Büring 2015) predicts pitch accents by combining a standard notion of focus (Rooth 1992), representing pragmatically 'new' information, with a set of rules that determine the *metric values* of nodes in a syntactic tree, representing the default stress patterns of a language. In this paper, we try to test whether this combination of focus and metric values indeed enables better prediction of prosodic stress, on a large corpus of spoken English. We do this by training standard machine learning models on predicting an automatically extracted measure of 'prosodic prominence' – a combination of pitch, duration and loudness – based on automatically extracted, approximate notions of pragmatic focus and syntactic structure (for calculating metric values). We find that, indeed, both sources of information contribute to predicting prosodic prominence, though with important caveats and open questions for future work.

Testing linguistic theories on naturalistic data is important (e.g., Chafe 1992) but remains challenging. In the case of focus semantics, no current corpora provide detailed information about the three dimensions involved: pitch accents, focus, and syntax. In this study, we rely on automatically extracted approximations of these notions.[1] The paper invites reflection on the use of such approximations for bridging the gap between theories and naturalistic data.

2 Background

2.1 Unalternative Semantics

As characterized in Büring 2015, standard focus theory (e.g., Rooth 1992) has three components. *Focus pragmatics* states that the focal target of a sentence (given the discourse context) must be among the set of focus alternatives. *Focus semantics* describes the relation between a sentence and its set of focus alternatives, standardly through the placement of F-markers on the syntactic tree. *Focus realization*, finally, concerns the relation between such F-markers and prosody, stating for instance that the main stress of a sentence must be placed within an F-marked constituent.

[1] Code for this paper can be found on https://github.com/fennaberger/prosodic-prominence.

Büring's Unalternative Semantics merges the latter two components – focus semantics and focus realization – by directly relating focus alternatives to prosody.

Unalternative Semantics relies on default rules about which words receive stress. One of these rules, called DEFAULT STRESS, is that, in a syntactic constituency tree, the right daughter of a node be prosodically stronger than the left daughter:

(1) DEFAULT STRESS
Label the left sister w(eak) and the right sister s(trong)

If this rule were the only rule in existence, the stress of a sentence would always be on the final word. This is called 'default prosody', and in such cases prosody and focus are related by the rule WEAK RESTRICTION:

(2) WEAK RESTRICTION (only applies in the case of default prosody)
If the weak daughter contains a constituent to be focused, the strong daughter does too.

This rule constrains which constituents may be the focus in case of default prosody. Consider example (3), which indeed shows default prosody:

(3) Mom made MUFFINS.

The rule WEAK RESTRICTION implies that 1) if "mom" is focused, the VP "made muffins" is focused too and 2) if "made" is focused, "muffins" is too. Hence, either the whole sentence is the focus, or the whole VP, or only the object DP – but not, for instance, only the subject DP.

If, contrary to the rule DEFAULT STRESS, a sentence contains a pitch accent somewhere on the left – so-called *prosodic reversal* – the rule STRONG RESTRICTION applies instead of WEAK RESTRICTION:

(4) STRONG RESTRICTION (only applies in the case of prosodic reversal)
The strong daughter contains a constituent to be focused, the weak daughter does not.

For example, (5) shows prosodic reversal at the S (sentence) node, since its left daughter is stronger than its right:

(5) DAD made pancakes.

Since DEFAULT STRESS cannot apply here, the rule STRONG RESTRICTION must be used, which implies that only "dad" can be the focus, nothing else.

At least two other rules (Büring 2016) play a role in determining default prosody (and, hence, co-define what counts as 'prosodic reversal'): "predicates (head) are weaker than their arguments (complements)" and "functional sisters (pronouns, auxiliars and similar) are weaker than lexical, open-class ones". These two rules will also be considered in the present study, along with DEFAULT STRESS, STRONG RESTRICTION and WEAK RESTRICTION.

2.2 The Helsinki Prosody Corpus

We seek to test the theory of Unalternative Semantics on naturally occurring spoken English. We will use the Helsinki Prosody Corpus (Talman et al. 2019), which consists of automatically generated annotations of 'prosodic prominence' on top of the LibriTTS corpus of recorded speech (Zen et al. 2019), using the Continuous Wavelet Transform (Suni et al. 2017) which measures fundamental frequency, energy, and duration in order to recognize prosodic events in speech files. They report an accuracy of 84% in detecting prosodic prominence. We use the 'train-360' partition of the Helsinki Prosody Corpus, containing around 2M words (see Table 1). In this corpus, each word has a *continuous* prominence score based on its prosodic prominence and a

Speakers	Sentences	Words	Prominence 0	Prominence 1	Prominence 2
904	116,262	2,076,289	1,003,454	569,769	503,069

Table 1: The train-360 partition of the Helsinki Prosody Corpus.

	Prosodic Prominence: 1 or 2	Prosodic Prominence: 0
Pitch Accent	737	99
No Pitch Accent	122	639

Table 2: Pitch accents and prosodic prominence

discrete prominence score of 0, 1, or 2, obtained simply by discretizing the continuous score. The distribution of prominence is skewed towards non-prominent, with 1M words receiving score 0, 569K words score 1 and 503K words score 2.

3 Method

One way to to test a theory of focus on naturalistic data, is to assess the ability of standard machine-learning models to predict pitch accent placement (the dependent variable), on the basis of features which the given focus theory deems necessary (the independent variables). For Unalternative Semantics, the features would include the focus of an utterance and the metric values for each syntactic node. As no corpus exists that contains all the necessary notions, we must rely on approximations: using prosodic prominence in place of pitch accents, using unpredictability in place of focus, and using automatic syntactic parsing for implementing Büring's rules for assigning metrical values (relative prosodic strength). These three approximations will be explained and discussed next, followed by the details of model fitting.

3.1 Prosodic Prominence as a Proxy for Pitch Accents

Whereas Büring's theory pertains to pitch accents, the Helsinki Prosody Corpus contains only automatic annotations of prosodic prominence – and not every word with high prosodic prominence will necessarily be perceived by human listeners as having a pitch accent. To assess the quality of this approximation, we annotate the pitch accents of 100 random sentences in the corpus, for a total of 1597 words. This was done by one rater (first author) who listened to the audio while using Praat (**broersma2019praat**) to visualize pitch contours. The results are shown in Table 2 (we group prominence scores 1 and 2 together, since both received a pitch accent in most cases).

According to these results, prominence scores are an accurate predictor of pitch accents 86% of the time. Recall that Suni et al. 2017 reported 84% accuracy for their method of automatically annotating prominence, although they did not look specifically at pitch accents. It is possible that the 86% accuracy we found is a compound of 1. inaccuracies in Suni et al.'s prosodic prominence scores to begin with, and 2. the difference between prosodic prominence as a graded notion and pitch accents as a discrete linguistic category.

3.2 Unpredictability as a Proxy for Focus

We will use unpredictability as an approximation of focus: words that are focused tend to be less predictable, more informative. Indeed, the notion of focus has often been characterized as the 'new' information, compared to what is 'given' (e.g., Schwarzschild 1999; Birner 1994), where 'given' does not simply mean to have been literally mentioned before, but, rather, to be unexpected given the context. We leave a quantitative assessment of the correlation between

unpredictability and focus for another occasion, and future work should try to tell the two apart (in particular because predictability is already known to correlate with prosodic prominence, Aylett and Turk 2004). (For the results reported here, we did not filter out the many tokens that were split up by BERT's word-piece tokenization, and used the first word-piece's probability.)

We implement this idea with the BERT language model (Kenton and Toutanova 2019), using the `transformers` library (Wolf et al. 2020). For each sentence, and for each word in that sentence, we created a version of the sentence where that word was masked (i.e., replaced by the special token `[MASK]`). We added one sentence before and one sentence after the sentence to provide context, and obtained the probability, according to BERT, of the original word being in the masked position. We used this probability, as a value between 0 and 1, directly as a measure of predictability, hence, as an approximation for *non*-focus. In order not to fool ourselves, we will henceforth use the term predictability instead of the notion of (non-)focus for which it is meant to be an approximation.

3.3 Automatically assigning metric values

We implement three rules that play a role in determining default prosody according to Büring 2016: "right sister is stronger than left sister" (DEFAULT STRESS), "complements are stronger than heads", and "content words are stronger than function words".

We implement the first rule ("right sister is stronger than left sister") using Stanza's constituency parser (Qi et al, 2020), by computing how often one has to turn 'right' to reach a given leaf node. However, whereas Büring's rule assumes binary trees, nodes in the constituency trees from Stanza can have more than two children (and do so in a variety of syntactic contexts). To deal with such cases, we assigned weights to each child corresponding to 'how far right' among the children they were. More precisely, we the score S_c of each child node, given its parent's score S_p can be calculated as follows, where n is the number of siblings the node has to its left, m is the total number of siblings (include the node itself), and $S_{root} = 0$:

$$S_c = S_p + \frac{n}{m-1} \tag{1}$$

We subsequently scaled all word-scores for a given sentence to the interval [0, 1] to control for sentence length (otherwise scores in longer sentences would be systematically higher).

To implement the second rule, "complements are stronger than heads", because Stanza's constituency parser does not provide information about heads and complements, we used SpaCy's (Honnibal et al. 2020) dependency parser instead (the model `en_core_web_md`; Unlabelled Attachment Score = 0.92, Labelled Attachment Score = 0.90). In general, the head of the non-head child of node X in a constituency tree depends in the dependency tree on the head child of node X (Xia and Palmer 2001), hence complements tend to be below their heads in the dependency tree. We thus implement the rule "complements are stronger than heads" by obtaining the depth of a word in the dependency tree, again scaled to control for sentence length.

Lastly, we implemented the rule "content words are stronger than function words" using SpaCy's part-of-speech tagger (`en_core_web_md`, POS accuracy 0.97), treating adjective, adverb, interjection, noun, proper noun, and verb as content words, and the rest as function words.

3.4 Models

We seek to evaluate Unalternative Semantics by assessing whether combining focus and metric values enables models to better predict pitch accents – or rather, given our approximations, whether combining predictability with metric values enables models to better predict prosodic prominence. We fit models in the 'predictability only' condition, the 'metric values only'

Total tokens	Total words (no punctuation)	Train set	Test set
100,000	70,667	53,000	17,667

Table 3: The train set and test set.

Type	Features	Accuracy	Precision	Recall	F1-score
Decision tree	Predictability	0.533	0.496	0.533	0.500
	Metric values	0.557	0.537	0.557	0.543
	Both	0.571	0.549	0.571	0.555
Random forest	Predictability	0.534	0.494	0.534	0.498
	Metric values	0.564	0.548	0.564	0.554
	Both	**0.579**	**0.557**	**0.579**	**0.563**
SVM	Predictability	0.521	0.409	0.522	0.457
	Metric values	0.566	0.551	0.566	0.556
	Both	0.568	0.547	0.568	0.553
Baseline	random	0.361	0.361	0.361	0.361

Table 4: Results of the classifiers (weighted average over classes), and weighted random baseline.

condition, and the combined 'predictability and metric values' condition, where the latter most closely resembles Büring's theory of Unalternative Semantics.

For each condition, we use `scikit-learn` to fit three types of classification models, to predict the discrete prominence scores (decision tree, random forest, and support vector machine (SVM)), and three types of regression models to predict the continuous prominence scores (ridge regression, random forest, and SVM).

Input features (predictability, metric values) were computed for an arbitrary total of 100K words, which after removing punctuation left 71K words for training and testing. We used the same train/test split for all six model types in all three conditions, using 25% as test set, and 5-fold cross-validation on the training set (see Table 3).

4 Results

See Table 4 for the results of the classification models and a weighted random baseline. For the decision tree and the random forest, the "both" (predictability and metric values) condition outperformed the other two conditions in all four metrics (all differences reported here are significant with $p < 0.0001$, according to Bowker-McNemar test with 3 dof). Not so for the SVM, where the "metric values only" condition in fact performed *very* slightly better than "both" in terms of F1-score (though not on accuracy and recall). As one possible explanation, we checked whether the SVM model in the "both" condition was perhaps overfitting on the training data, but this was not the case; the model did not perform better on the training data.

We inspect the best classification model more closely, the random forest in the "both" condition, by looking at the feature importances. The rule "content words are stronger than function words" carries the most weight (0.48), followed by predictability (0.37), "right sisters stronger than left sisters" (0.1) and "complements stronger than heads" (0.05).

The confusion table for our best classifier revealed substantial confusion between levels 1 and 2 (i.e., medium and strong prominence). Our annotations earlier revealed that both levels tend to reflect pitch accents, suggesting we could merge these levels, turning the task into a binary classification task. Moreover, this meant we could take our 86% accuracy of the prominence-as-pitch-accents approximation into account, by randomly reassigning labels in the test set according to this probability. Average F1-scores over 100 thusly resampled test sets revealed

Type	Features	Mean	Standard deviation
Ridge	Predictability	0.556	0.461
	Metric values	0.541	0.455
	Both	0.514	0.452
Random forest	Predictability	0.550	0.459
	Metric values	0.533	0.451
	Both	0.503	0.448
SVM	Predictability	0.519	0.539
	Metric values	0.499	0.531
	Both	**0.480**	**0.512**
Baseline	Mean	0.634	0.462

Table 5: Results (mean error and standard deviation) for the regression models (graded prominence).

that the best model was still the random forest in the "both" condition (mean F1-score of 0.690), marginally better than the "metric values only" condition (0.684) and more substantially so for the "predictability only" condition (0.640), and in each case significantly so. By contrast the decision tree proded significantly superior in the "metric values only" condition (0.681) compared to "both" (0.675) and "predictability only" (0.629), while no significant difference was found for the SVM classifier.

Table 5 shows the results of our regression models predicting graded prosodic prominence scores (between 0 and 2), compared to the average as a baseline. We report mean difference between predictions and targets. For all three types of regression models, the "both" condition had a better fit than the conditions with either only predictability or only metric values (all comparisons significantly so as per T-test, $p < 0.01$).

We inspect the best regressor more closely, i.e., the SVM model in the "both" condition, by estimating feature importances by permutation (Breiman 2001). These show the same pattern as the random forest classifier above: "content words are stronger than function words" carries most weight (0.244), followed by predictability (0.128), then "right sisters are stronger than left sisters" (0.04) and "complements are stronger than heads" (0.003).

The Ridge regressor is our only model which computes a likelihood function, so we can compute its Akaike information criterion (AIC) to assess whether the improved fit in the "both" condition is worth added model complexity due to it having more features. This appears to be the case: the AIC score in the "both" condition (111701) is significantly lower than the scores of the "predictability only" condition (119251) and the "metric values only" condition (116221).

5 Discussion

We conclude that having both predictability and metric values helps models predict prosodic prominence, as opposed to having only one of the two. Moreover, in the resulting models the content/function word distinction and predictability carry most of the weight. On the surface, none of the models seem particularly successful (top F1-score of 0.56, or 0.69 in the binary condition). Presumably this reflects, at least in part, inaccuracies in our approximations (unpredictability as focus, prominence as pitch accents, imperfect syntactic parsing). In addition, the default prosody rules identified by Büring may not be the only ones. However, ultimately, accent placement is determined not by context, but by a speaker's communicative intentions. For a more definitive assessment of the quality of model fit, it would be helpful to compare it to the ability of *humans* to predict prominence based only on the transcripts.

References

Aylett, Matthew and Alice Turk (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech". In: *Language and speech* 47.1, pp. 31–56.

Birner, Betty J (1994). "Information status and word order: An analysis of English inversion". In: *Language*, pp. 233–259.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Büring, Daniel (2015). "Unalternative semantics". In: *Semantics and linguistic theory*, pp. 550–575.

— (2016). "A beginner's guide to unalternative semantics". In: *Manuscript University of Vienna. http://homepage. univie. ac. at/daniel. buring/phpsite/index. php*.

Chafe, Wallace (1992). *The importance of corpus linguistics to understanding the nature of language*. De Gruyter.

Honnibal, Matthew et al. (2020). "spaCy: Industrial-strength natural language processing in Python". In.

Kenton, Jacob Devlin Ming-Wei Chang and Lee Kristina Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of naacL-HLT*. Vol. 1. Minneapolis, Minnesota, p. 2.

Rooth, Mats (1992). "A theory of focus interpretation". In: *Natural language semantics* 1.1, pp. 75–116.

Schwarzschild, Roger (1999). "GIVENness, AvoidF and other constraints on the placement of accent". In: *Natural language semantics* 7.2, pp. 141–177.

Suni, Antti et al. (2017). "Hierarchical representation and estimation of prosody using continuous wavelet transform". In: *Computer Speech & Language* 45, pp. 123–136.

Talman, Aarne et al. (2019). "Predicting prosodic prominence from text with pre-trained contextualized word representations". In: *arXiv preprint arXiv:1908.02262*.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45.

Xia, Fei and Martha Palmer (2001). "Converting dependency structures to phrase structures". In: *Proceedings of the first international conference on Human language technology research*.

Zen, Heiga et al. (2019). "LibriTTS: A corpus derived from librispeech for text-to-speech". In: *arXiv preprint arXiv:1904.02882*.